NOVEL NON-PARAMETRIC METHODS FOR UNCOVERING STRUCTURE IN DATA: ESTIMATION OF PLACE AND GRID CELL FIELDS

by

Rahul Agarwal

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy (PhD)

Baltimore, Maryland

August, 2015

Abstract

Maximum likelihood (ML) estimators of probability density functions (pdfs) are the most popular parametric estimators today because they are often efficient to compute and have several nice properties such as consistency, asymptotic normality, functional invariance, achieving a lower bound on the variance of the estimator parameters (Cramer-Rao bound) and they have fast convergence rates. However, often the underlying data is too complex and it is not easy to parametrize the pdf. In such cases, non-parametric modeling remains the only option.

Existing non-parametric methods, such as kernel density estimation (KDE), orthogonal series density estimates (OSDE) and orthogonal series square-root density estimates (OSSDE) are consistent. However, these estimators do not necessarily have other properties of parametric ML estimators and have slower convergence rates. On the other hand, non-parametric ML estimation has not been rigorously studied, because in general the likelihood is hard to maximize in a non-parametric setting. One example of a non-parametric ML estimator is the histogram (or the experimental cumulative distribution function). The histogram is a consistent estimator, but it is discontinuous. Many pdfs in nature are smooth and hence it is desirable to obtain smooth estimators.

This thesis proposes a nonparametric ML estimator over the set of band-limited (BL) "smooth" pdfs - the BLML estimator. This class contains pdfs whose Fourier transforms have finite support (with a certain cut-off frequency). A semi-closed form of the BLML estimator is derived and its consistency is shown. Although convergence rates are not derived, the BLML estimator has faster convergence rates than KDE and OSSDE methods in simulation. Algorithms for fast computation of the BLML estimators are also proposed and their computational complexity is determined to be better than that of KDE and OSSDE methods. In fact, in simulation these BLML algorithms show an order of magnitude faster computational time than the KDE and OSSDE methods for dense data. Finally, algorithms for estimating the unknown cut-off frequency are proposed.

BLML methods are then used to construct an estimator for mutual dependence between different random variables. Mutual dependence measures the Bhattacharya distance between the joint pdf and the product of marginal pdfs and is an "ideal" metric for measuring dependencies between random variables unlike measures such as the mutual information, Pearson or distance correlation. Currently mutual dependence is not directly estimable from data and its estimation requires numerical integration which can produce errors.

The consistency of the BLML estimator for mutual dependence is then proven and simulations are used to show that the convergence rates of the BLML estimator for mutual dependence are superior to the convergence rates of the OSSD estimator for mutual dependence, the estimator for Pearson and distance correlation. Finally, an algorithm to estimate the cut-off frequencies for the BLML estimator for mutual dependence is developed and its

performance on standard dataset from Wikipedia is shown.

Then BLML methods are then applied to estimate the encoding fields of complex grid and place cells. Recently introduced "Fourier hypothesis" states that grid cell fields are approximately 2-dimensional cosine fields and place cell fields are the linear sum of such grid cell fields (1). Due to the close relation of the BLML methods to Fourier transforms, BLML estimation is a natural choice for testing this hypothesis. In particular, the conditional intensity function of 53 place and grid cells is estimated using the BLML methods and the Bayesian framework. The performance of the BLML methods is then compared with that of KDE and generalized linear models (GLM) (which are state-of-the-art parametric methods in neuroscience). The BLML methods outperform both the KDE and GLM methods validating the hypothesis.

Further, the BLML (along with KDE) methods also successfully captures the history dependence in firing patterns of place and grid cells. Thereby, these methods are able to explain the variance that has previously been observed in firing patterns of place cells (2) and has not been explained by other models for complex place fields. Finally, the BLML methods are used to decode the trajectory of rat (using the firing patterns of the grid and place cells) with considerable accuracy ($r^2 = 0.89$).

Thesis advisor:

Sridevi V. Sarma

Assistant Professor

Department of Biomedical Engineering

Johns Hopkins University

Thesis committee:

Dr. Steve G. Massaquoi

Dr. Nitish V. Thakor

Dr. Rene Vidal

To my family.

with love Rahul

Acknowledgements

This thesis is done as a part of Doctor of Philosophy (PhD) requirement at the department of Biomedical Engineering, Johns Hopkins University. The quality of the work has been improved manifolds due to the support, teachings and technical expertise I recieved from my mentor, collegues and friends. I am acknowledging their contributions, heretoforth. I would first like to acknowledge my advisor Dr. Sridevi Sarma. It is due to her support and advice that this work is made possible. During my stay of six years in the lab as a MS and PhD student, I moved across a lot off different projects and I am grateful to Sri that she readily mentored me even though some of the projects were not her expertise. Apart from technical mentoring, I am also grateful to Sri for her support and belief in me as a good researcher. Her calm and soothing personality helped me remain stable during the roller coaster journey of the PhD. Second, I would like to thank Dr. Steve Massaquoi for his valuable support and advise during a project that entailed understanding the neuronal circuitry in the pre-motor cortex. Although, this project did not end up as a part of my thesis but it is a valuable contribution from my research. I learned a lot by working with Steve, particularly the importance of perseverance and developed an eye for details. Third, I would like to thank my thesis committee members Dr. Rene Vidal and Dr. Nitish Thakor,

for their valuable feedback on my thesis work. I am very grateful that both the members made time for discussing my research whenever I needed. These meetings resulted in very valuable feedback which has significantly improved the quality of this thesis. Fourth, I would like to thank Dr. John Gale, Dr. Marc Schieber and Dr. Zhe (Sage) Chen for helping me with neuroscience experiments and data collection. If not for the data made available by Dr. Chen, Dr. Gale and Dr. Schieber, I would not have came across problems that are faced by neuroscientists while analyzing such data. These problems later formed the basis for this thesis. Fifth, I would like to thank Dr. Narendra Dixit and Dr. Emery Brown who helped me to start my research career by allowing me to work in their labs as an undergraduate summer intern. This developed my interest in research and made me pursue graduate studies. Sixth, I would like to thank Dr. Pierre Sacre, Dr. Munther Dahleh, Dr. Sanjeev Khudanpur, Dr. Mesrob Ohannessian and Dr. Ben Haeffale for doing valuable discussions on the BLML estimator. These discussions helped a lot in shaping this thesis. Finally, I would like to thank my parents, brother, brother in-law, sister, sister-in-law and friends who were always a source of emotional support and inspiration for me. This thesis was not possible without their support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction: A Motivating Example

The goal of modeling in data science is to describe a random variable of interest as a function of other variables, called "covariates," from measurable data. The functional relationship is formalized by computing an estimate of the joint probability density function (pdf) between all random variables. Several challenges may arise when estimating unknown pdfs from data. First, it may be difficult to compute an estimate from data, especially in cases where no structure is visible in the data. Second, even if an estimate is computable, it may not be straightforward to show that it converges to the true pdf as the number of samples goes to infinity (a notion called "consistency"). Third, if the estimate is consistent, it may be difficult to derive the rate of convergence which is important to understand how much data is required for a reasonable estimate. Finally, the number of covariates may be very large, *the curse of dimensionality*, and dimensionality reduction may be required.

Reducing dimensionality entails identifying statistical dependencies between variables, and then eliminating variables that carry redundant information.

Modeling from data is often required in neuroscience wherein in a functional relationship between a neural response to an applied stimulus or to an induced behavior is needed. For example, a rat is placed in a circular arena and motivated to move freely about (figure 1.1A), while micro-electrodes are implanted into the hippocampus and the entorhinal cortex, two regions of the brain that encode the rat's position. The micro-electrodes record spiking activity of multiple neurons, which can then be post-processed to identify the activity of the individual neurons - a process called spike sorting (3). While the neural activity is recorded, the rat's position is simultaneously measured by placing two infra-red diodes alternating at 60 Hz attached to the micro-electrode array drive implanted in the animal. The recorded trajectory of rat and firing pattern of three sample cells are shown in figure 1.1B,C,D. The goal of such an experiment is to understand how the neurons encode position information and if they encode anything else about the rat's behavior or the environment.

Such navigation experiments have shown that neurons in the entorhinal cortex and hippocampus exhibit complex "grid-like" or "place-like" spiking patterns as a function of rat's position (4, 5) (see figure 1.1). However, it is unclear if variables other than rat's positional co-ordinates also influence spiking activity of each neuron. In particular, entorhinal cortex and its closely related hippocampus are also known to be involved in memory mechanisms in the brain, therefore the dependence of grid and place cell activity on their spiking histories (denoted as H) is possible. To understand such relationships, it is required to model the

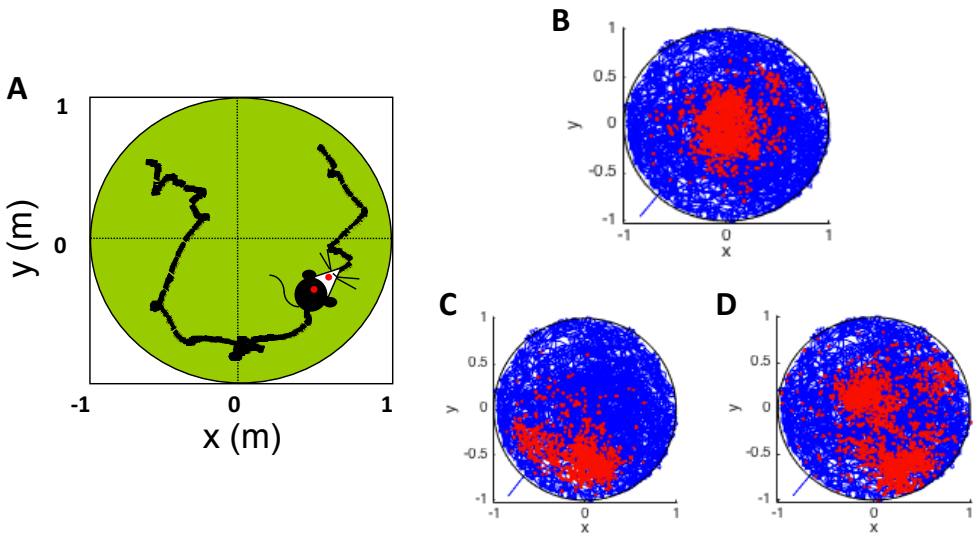

Figure 1.1: Spiking activity of grid and place cells - **(A)** A schematic of the circular arena where the rat was freely foraging. **(B,C,D)** Spiking activity of a simple place cell, a complex place cell and a grid cell respectively. The trajectory of rat foraging freely in the circular arena (for $40min$) is marked by blue lines. The x, y co-ordinates of a rat's position when the grid cell spiked are marked by red dots.

neuronal spiking activity of these cells as a function of the rat's position and the neuron's own spiking history.

The spiking activity of a neuron can be modeled as a stochastic *point process* which is a series of 0-1 random events that occur in continuous time (6, 7, 8). For a neural spike train, the 1-s are individual spike times and the 0-s are the times at which no spikes occur. To define a point process model of neural spiking activity, consider an observation interval $(0, T]$, and let $N(t)$ be the number of spikes counted in interval $(0, t]$ for $t \in (0, T]$. A point process model of a neural spike train can then be completely characterized by its conditional

intensity function (CIF), $\lambda(t|\mathbf{x})$, defined as follows:

$$\lambda(t|\mathbf{x}) = \lim_{\Delta \to 0} \frac{Pr(N(t+\Delta) - N(t) = 1|\mathbf{x})}{\Delta}. \tag{1.1}$$

where $\mathbf{x}$ denotes a vector of covariates that can contain history of spiking. In other words, the CIF for point process generalizes the rate function of an inhomogeneous Poisson process to a rate function that can also be dependent on spiking history. Now, it follows from (2) that the probability of a single spike in a small interval $(t, t+\Delta]$ is approximately

$$Pr(\text{spike in } (t, t+\Delta] \ |\mathbf{x}) \ = \lambda(t|\mathbf{x})\Delta \tag{1.2}$$

Details can be found in (6, 9). When Δ is small, (2) is approximately the spiking propensity at time t. Therefore, the CIF completely characterizes a spike train and hence a validated model for the CIF as a function of covariates can reveal the dependence of the neuronal spiking on the covariates. A popular method for modeling the CIF is by using generalized linear models (GLMs) (10, 11), which assume that the CIF is a log-linear in the parameters:

$$log(\lambda) = \sum_i \alpha_i \phi_i(x_i). \tag{1.3}$$

Here, $\phi_i(x_i)$s are the covariates which are functions of measured variables in the experiment, and the α_i's are the parameters of the model. For, point processes the GLM assumption yields a convex likelihood function that can be efficiently maximized for estimating the parameters (11, 12). Additionally, the functions $\phi_i()$s can be chosen to capture

more complex dependencies of the CIF on measured variables. This framework has been successful in describing CIF in neurons in various regions of the brain, e.g. Basal Ganglia, Hippocampal place cells etc (13, 14, 15, 16, 17, 18, 19, 20). However, it has limitations in describing more complex firing patterns such as those shown in figure 1.1 C,D. Recently Kloosterman et. al. (21) suggested the use of non-parametric methods, particularly kernel density estimation under a Bayesian framework. In this approach, Baye's rule is used to describe CIF as a ratio of probability density function (pdfs) as follows:

$$\lambda(t) \simeq \frac{N}{T} \frac{f(\mathbf{x}|\text{spike in time } \Delta t)}{f(\mathbf{x})}, \tag{1.4}$$

where N is the total number of spikes within the total duration of the spike train observation T - $N(T)$. $f(\mathbf{x})$ and $f(\mathbf{x}|\text{spike in time } \Delta t)$ are probability densities which can be estimated parametrically or non-parametrically.

More recently Ormond et. al. (1, 22) hypothesized that the place fields are formed as a result of Fourier like summation of grid cell fields (see figure 1.2). The grid cell field can closlely be approximated by 2-dimensional cosine function with some spatial frequency. These cells are generally arranged in anatomical regions called modules (see figure 1.2 A) that contains the grid cells that have similar saptial frequencies (23). The place cells in hippocampus gets projections from grid cells across multiple modules (having different spatial frequencies), which in turn act as a summer of the grid cell fields with different frequencies. A toy example of such a sum is shown in 1.2B. This hypothesis is called the "Fourier hypothesis". If the Fourier hypothesis is true the estimates of densities that assume

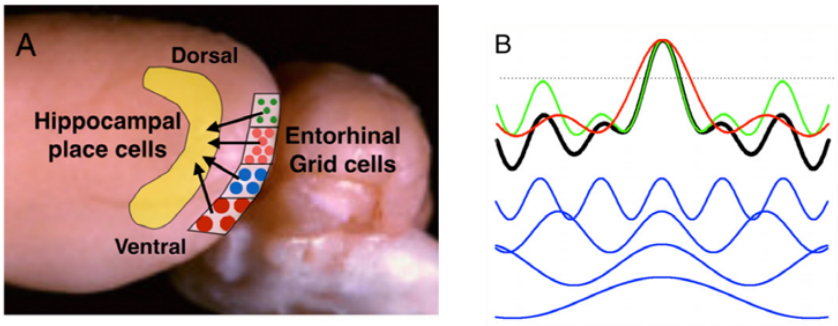

Figure 1.2: The Fourier hypothesis - **(A)** Lateral view of entorhinal cortex and hippocampus. The grid cells are arranged in modules within entorhinal cortex. Each module contains grid cell that have fields with similar spatial frequency. The place cells in hippocampus recieves synaptic projections from grid cells in different modules. The anatomical connections provides evidence that the place cell may linearly sum the grid fields to generate a place field as shown in **(B)**. The figure is taken from (22).

that the place and grid fields are supported by sinusoidal ad/or cosine functions as in fourier representation may be able to capture the grid and place fields more accurately.

In addition to estimating point process pdfs of neuronal spike trains, it is also important to identify statistical dependencies between different covariates to ultimately eliminate covariates that carry redundant information. For instance, the rat's position at a given time t, $(x(t), y(t))$ may be statistically dependent on $H(t)$, the spiking history up to time t. What may be tricky is understanding which history covariates defined on $x(t), y(t)$ are dependent. Density and dependency estimation are discussed in next two sections.

1.1 Density estimation

In general, the problem of density estimation (24) entails estimating a pdf $\hat{f}(x; x_1, \cdots, x_n)$ that is a function of $i.i.d$ observations $x_1, \cdots, x_n$ of a random variable X with true pdf $f(x)$. Ideally, $\hat{f}(x)$ should be close to $f(x)$, in some sense (e.g. L_2 error), it should be unbiased i.e. $\mathrm{E}(\hat{f}(x)) = f(x)$, it should be easy to compute from data, and it should have minimum variance over all possible estimators $\hat{f}(x)$. Finally, the estimator $\hat{f}(x)$ should be consistent, which means that $\hat{f}(x)$ should converge to true density $f(x)$ as the number of samples, n, go to infinity.

Finding an estimator that satisfies the aforementioned properties is in general difficult. However, in the parametric setting, where it is assumed that the true density lies in some class of functions parametrized by a vector θ, i.e.,

$$f(x) = f(x; \theta), \tag{1.5}$$

these properties can be achieved by maximizing the data likelihood function over θ:

$$\hat{\theta}_{MLE} = \mathrm{argmax}_\theta(\mathcal{L}(\theta; x_1, \cdots, x_n)), \tag{1.6}$$

where $\mathcal{L}(\theta; x_1, \cdots, x_n) \triangleq \Pi_{i=1}^n f(x_i; \theta)$ is the likelihood function of observing $x_1, \cdots, x_n$ for a given parameter value θ. Such estimators are called *parametric* maximum likelihood (ML) estimators and are often efficient to compute. However, if the true density does not lie

in the assumed class of parametric functions, the ML estimates fail to achieve the desirable properties. In such cases, non-parametric estimation has to be used.

The histogram is the most fundamental non-parametric density estimator. It estimates the density by dividing the domain of the unknown pdf into equal size bins (intervals) and assigns p_j as the Bernoulli probability (10, 25) for observing a sample in jth bin. A trivial estimator (also ML) for p_j is $\hat{p}_j = \frac{1}{n}\sum_{i=1}^{n} I_j(x_i)$ where $I_j()$ is the indicator function for bin j. The estimator for the pdf then is $\hat{f}(x) = p_j/A$ if $x \in \text{bin}_j$ where A is the area/volume of the bins.

In general, non-parametric ML estimators are computed as follows:

$$\hat{f}(x; x_1, \cdots, x_n) = \text{argmax}_{f \in \mathcal{F}}(\mathcal{L}[f; x_1, \cdots, x_n]), \tag{1.7}$$

where $L[f; x_1, \cdots, x_n] = \Pi_{i=1}^{n} f(x_i)$ is the likelihood functional for observing the data $x_1, \cdots, x_n$, and $\mathcal{F}$ is a non-parametric class of functions. In this thesis, the term non-parametric estimation is used when the number of parameters needed to define the class $\mathcal{F}$ is unbounded. For histograms, this class is the linear span of shifted rectangular basis functions (bins), and since $\hat{f}(x) = 0$ in the bins which do not contain any data samples, histogram estimators are tractable and easy to compute. However, shifted rectangular basis functions results in estimates of $f(x)$ which are non-smooth, piecewise constant, and that depend on the choice of bin size and bin center locations. Further, the histogram estimator is consistent only if the bin-width goes to zero as the number of samples increases. Finally,

the advantage of tractability and computability is generally lost in high dimensional data sets, since the number of plausible bins and centres grows combinatorially.

In general, it is difficult to maximize likelihood over any given non-parametric class. In such cases, penalized likelihood methods are used, which involves adding a penalty term to the likelihood reward function (a procedure also known as "regularization" (26, 27)). Although regularization allows one to compute the solution to the penalized likelihood function, the nice properties of ML methods may be lost.

Other non-parametric estimators, such as kernel density, orthogonal series density, orthogonal series square-root density, and K-nearest neighbor, focus mainly yielding smooth, non-negative consistent estimates with tolerable convergence rates, but they do not maximize likelihood. Also the non-negativity of the estimates is generally traded for achieving higher convergence rates (a detailed literature survey of these methods is done in chapter 2).

In summary, none of the current density estimators simultaneously have the following four desirable properties when structure is not apparent in measured data:

- Non-parametric: Enables density estimation without worrying about the class of functions in which the true pdf belongs.
- Maximum likelihood: Asymptotically may have several desirable properties for the estimator.

- Smooth: Many real densities are smooth. Therefore, a smooth estimator may represent such densities more accurately.
- Non-negative: Pdfs are by definition non-negative. Therefore, it is desirable to have a non-negative estimator.

This thesis focuses on finding a density estimator that simultaneously has the above desirable properties.

1.2 Dependency estimation

Dependency estimation entails quantifying the amount of information contained in one random variable about another random variable. The amount of shared information, in general, can be quantified using the joint density of the two variables (if known) as following:

$$d = ||f(x, y) - f(x)f(y)|| \tag{1.8}$$

where d is the metric for dependency, $f(x, y), f(x)$ and $f(y)$ is the joint and marginal densities respectively and $|| \cdot ||$ is some norm. However, in the case when the explicit representation of the joint density is unknown and only data samples from the joint density are available, estimating the dependency is challenging. Computing dependencies between random variables entails first selecting a theoretical measure for the density (having some desirable properties, as listed in table 1.1 and then estimating the measure directly from data.

Table 1.1: Desired properties of ideal dependency measure $\delta(X, Y)$.

#	Property	r	I	R	d
1	$\delta(X,Y) = \delta(Y,X)$	✓	✓	✓	✓
2	$\delta(X,Y) = 0$ iff X and Y are independent		✓	✓	✓
3	$0 \leq \delta(X,Y) \leq 1$			✓	✓
4	$\delta(X,Y) = 1$ if there is a strict dependence between X and Y			✓	✓
5	$\delta(X,Y) = f(r(X,Y))$ if the joint distribution of X and Y is normal	✓	✓	✓	✓
6	$\delta(\psi_1(X), \psi_2(Y)) = \delta(X,Y)$		✓		✓
7	$\delta(X,Y)$ is a metric			✓	✓

Note: ψ_1, ψ_2 are strictly monotonic functions

Due to the complexity of estimating the measure directly from data, popular dependency measures used today are Pearson's correlation, r, and distance correlation, R, (28, 29, 30) which do not satisfy all desirable properties. On the other hand, mutual dependence, d, (31) and mutual information I (32) do satisfy the above properties but are not estimable from data in any straightforward manner. Therefore, this thesis also focuses on finding the measure for dependencies that satisfies all seven desirable properties listed in table 1.1 and that is directly estimable from the data.

The dependency measures mentioned above and estimation approaches are discussed in detail in next chapter.

Chapter 2

Prior Art

This chapter describes current methods for estimating pdfs, dependencies between random variables, and the CIF of point processes.

2.1 Density estimation

There are two approaches for estimating pdfs: parametric and non-parametric, which are described below.

2.1.1 Parametric density estimation

Parametric density estimation assumes that the unknown pdf is a member of a known parametric class of functions, $f(x;\theta)$, parametrized by a vector θ. Some examples of such families are the exponential distribution family $f(x;\lambda) = \lambda e^{\lambda x}$ with parameter λ, the normal distribution family $f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$ with parameters μ and σ, and the uniform distribution family $f(x;a,b) = \frac{1}{b-a}\ if a < x \leq b$ and 0 otherwise, with parameters a, b (33).

Consider $x_1, \cdots, x_n$ independent and identically distributed (i.i.d.) experimental outcomes from a pdf $f(x;\theta_0)$ with true parameter θ_0. Then, an estimator $\hat{\theta}(x_1, \cdots, x_n)$ for θ_0 can be any function $\hat{\theta}$ of the observed data $x_1, \cdots, x_n$. Out of several possibilities, the estimator functions that satisfy certain properties are mostly preferred. Two such estimators are the minimum variance unbiased estimator and the maximum likelihood estimator described next.

2.1.1.1 The minimum variance unbiased parametric estimator (MVUE)

An unbiased estimator is defined as the estimator for which $\mathrm{E}(\hat{\theta}) = \theta_0$. The MVUE is an unbiased estimator that minimizes the variance of $\hat{\theta}$ (34), i.e.,

$$\hat{\theta}_{MVUE} = \mathrm{argmin}_{\hat{\theta}:\mathrm{E}(\hat{\theta})=\theta_0}(\mathrm{Var}(\hat{\theta})) \tag{2.1}$$

Note that the expected value is defined as $\mathrm{E}(\hat{\theta}) = \int \hat{\theta} f(x_1, \cdots, x_n; \theta)\mathrm{d}x$ and $\mathrm{Var}(\hat{\theta}) = \mathrm{E}(\hat{\theta}^2) - \mathrm{E}(\hat{\theta})^2$.

Finding the MVUE is not always easy as it may not exist and if it does exist (guaranteed by the existence on an unbiased estimator (35)), finding the MVUE requires finding a sufficient and complete statistic, which is difficult (36). Therefore in such cases other methods have to be used.

2.1.1.2 The maximum likelihood (ML) parametric estimator

ML estimation (37) is the most popular parametric estimation method. It maximizes the data likelihood function $\mathcal{L}(\theta; x_1, \cdots, x_n) \triangleq \Pi_{i=1}^{n} f(x_i; \theta)$, i.e.,

$$\hat{\theta}_{MLE} = \text{argmin}_{\theta}(\mathcal{L}(\theta; x_1, \cdots, x_n)) \tag{2.2}$$

ML estimators have several nice asymptotic properties (37) when the actual distribution lies in the assumed parametric class:

- Consistency- $\hat{\theta} \rightarrow \theta_0$ as number of samples $n \rightarrow \infty$
- Statistical Efficiency- ML estimators achieve the minimum variance over all unbiased estimators (Cramer-Rao lower bound) asymptotically and hence are MVUE asymptotically.
- Asymptotic Normality- $\sqrt{n}(\hat{\theta} - \theta_0) \sim N(0, I^{-1})$ asymptotically. Here $I \triangleq \text{E}\left(\frac{\text{d}^2 L(\theta)}{\text{d}\theta^2}\right)$ is the Fisher information matrix.
- Functional invariance- For the functional relationship $\alpha_0 = g(\theta_0)$, $\hat{\alpha}_{ML} = g(\hat{\theta}_{ML})$.
- Fast Convergence- In practice, for a ML estimator, the mean squared error (MSE) decreases as $\mathcal{O}(n^{-1})$ which is the fastest possible convergence rate for an estimator.
- Fast Computation- ML estimators are often easy and quick to compute if the likelihood function is convex in the parameters.

- Interpretability- The parameters may be related back to physiological and environmental variables, which is true for any parametric approach.

However, if the true pdf does not lie in the assumed class of functions, large errors may occur, potentially resulting in misleading inferences. Further, when the likelihood function is non-convex in parameters it may become computationally difficult to compute the ML estimator exactly. Therefore a need for non-parametric estimation methods arises.

2.1.2 Non-parametric density estimators

Parametric density estimation can only be used if the parametric density class in known a priori. More often than not such parametric density classes are unknown. In such cases non-parametric density estimation methods are very useful. Popular non-parametric density estimation methods are described next.

2.1.2.1 K nearest neighbors estimators

K nearest neighbors (KNN) (38) estimates eliminate the need of knowing bin center locations (needed for the histogram estimator) and estimate the unknown density at a point x as:

$$\hat{f}(x) = \frac{k}{nV_k(x)} \tag{2.3}$$

where $k \in \mathbb{N}$ governs smoothness of the estimator. $V_k(x)$ is the minimum volume of a sphere that is centred at x and encompasses kth nearest neighbor (in the observation set) of x, and n is the total number of observation points. KNN estimators have some

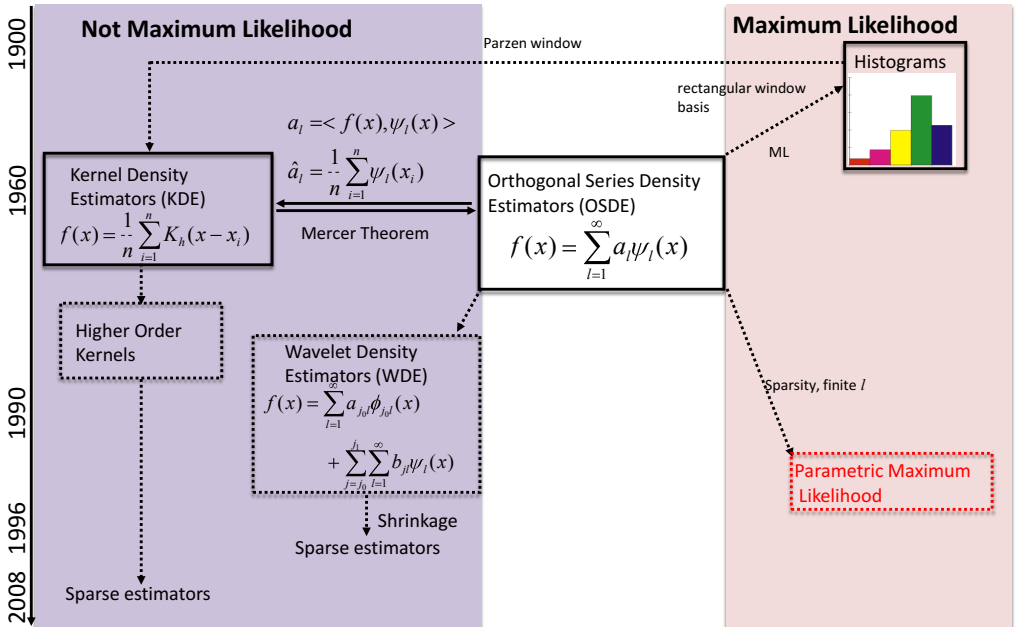

Figure 2.1: History of non-parametric methods - Showing historic evolution of non-parametric methods and the relations in between. As can be seen in the figure and explained in the text the popular non-parametric methods can be linked to orthogonal series density estimation.

computational advantages, but suffer from many drawbacks that render them of little use. These drawbacks include dependence on local noise, very heavy tails($\frac{1}{x}$), non-smoothness, and most importantly, KNN estimators are technically not pdfs as the integral of $\hat{f}(x)$ diverges.

2.1.2.2 Kernel density estimators

Kernel density estimation (KDE) (39, 40, 41, 42, 43) is the most widely used non-parametric method as it yields smooth estimates of pdfs and eliminates the dependence on bin locations.

KDE assumes that the density is a linear sum of kernel functions that are centred on the data points, i.e.

$$f(x) = \frac{1}{nh} \sum_i K(\frac{x - x_i}{h}). \tag{2.4}$$

In above equation, $h \in \mathbb{R}$, is called the "bandwidth" of the kernel and is akin to the bin width parameter in a histogram. Intuition for KDE comes from numerical approximation of a pdf by the empirical cumulative density function ($F_n(x) \triangleq \frac{1}{n} \sum_{i=1}^{n} I(x_i < x)$). Specifically, consider the following numerical approximation for the derivative of empirical CDF:

$$f(x) = \frac{F_n(x) - F_n(x - h)}{h} \tag{2.5}$$

$$= \frac{1}{nh} \sum I_{(x-h,x]}(x_i) \tag{2.6}$$

$$= \frac{1}{nh} \sum I_{(0,1]} \left(\frac{x - x_i}{h} \right), \tag{2.7}$$

which gives the KDE with rectangular kernels. Higher order numerical approximations of the derivative results in higher order kernels. In general, the order of a kernel (44), is defined as minimum $\nu \in \mathbb{N}$ such that

$$\int u^\nu K(u) \mathrm{d}u \neq 0. \tag{2.8}$$

It can be shown that all symmetric non-negative kernels are second order kernels, therefore higher order kernels lead to estimates of pdfs that can be negative for some values of x. However, error analysis has shown that higher order KDE achieves faster convergence

rates (44). In particular, it has been shown that for a ν order kernel, the mean integrated squared error (MISE) defined as:

$$\text{MISE} \triangleq \text{E}\left(\int \left(\hat{f}(x) - f(x)\right)^2 \text{d}x\right), \tag{2.9}$$

goes as $\mathcal{O}\left(\frac{1}{nh}\right) + \mathcal{O}\left(h^{2\nu}\right)$. The optimal convergence rate is then achieved if the bandwidth h goes as $\mathcal{O}\left(n^{\frac{2\nu}{2\nu+1}}\right)$, which results in a convergence rate of $\mathcal{O}\left(n^{\frac{1}{2\nu+1}}\right)$ for the MISE.

However, KDE does not maximize likelihood and needs knowledge of the bin width a priori which should again go to zero asymptotically to achieve consistency (this results in slower convergence rates). Further, choosing the kernel functions is also a tricky and often an arbitrary process (45) and have been under study for decades. Additionally, even the best KD estimators (45, 46, 47, 48) have slower convergence rates ($\mathcal{O}_p(n^{-4/5})$, $\mathcal{O}_p(n^{-12/13})$ for the second and sixth-order Gaussian kernels, respectively) than the parametric ML estimation ($\mathcal{O}_p(n^{-1})$) for the mean integrated squared error (MISE)(49).

2.1.2.3 Orthogonal series and orthogonal series square-root density estimators

Orthogonal series density estimation (OSDEs) (50, 51) is probably the second most popular non-parametric method. It is closely related to KDE, but is a bit more general as described next. OSDE assumes that the unknown density is in the linear span of an orthonormal series, i.e.,

$$f(x) = \sum_{j=1}^{\infty} a_j \psi_j(x). \tag{2.10}$$

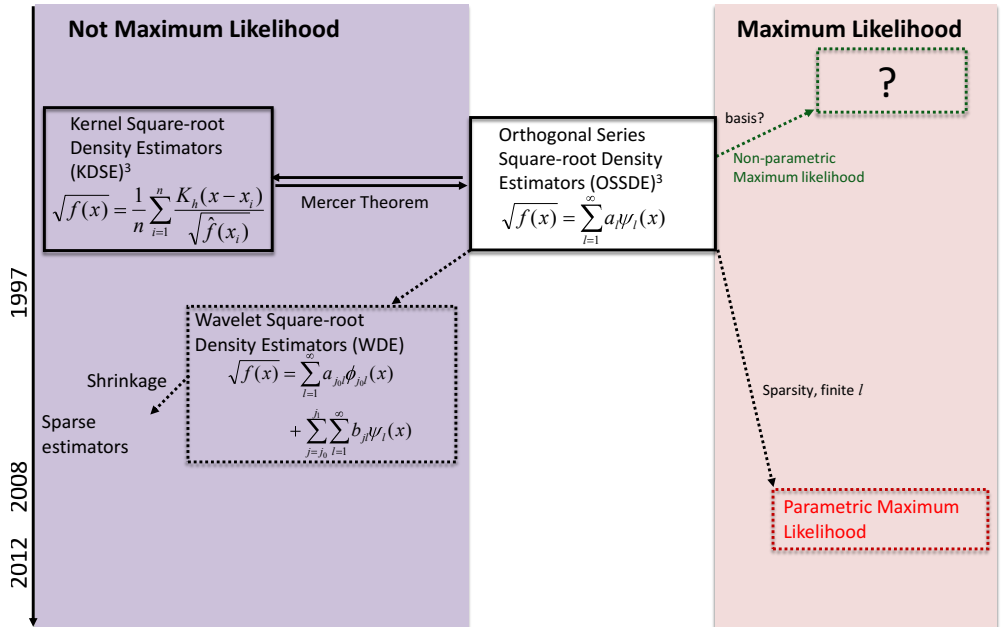

Figure 2.2: History of non-parametric methods in square-root domain - Showing historic evolution of non-parametric methods in square root domain and the relations in between. On comparing it with figure 2.1 it can be noticed that although, in square-root domain several corresponding estimators (to linear domain) has been discovered, there remains a glaring gap where an estimator that maximizes the likelihood non-parametrically is needed to be discovered.

The coefficients a_j can be estimated by one of the three methods outlined next. The first method uses the orthonormality of the basis, due to which

$$a_j =< \psi_j(x), f(x) >= \mathrm{E}(\psi_j(x)) \tag{2.11}$$

Therefore, each a_j can be approximated as $\hat{a}_j = \frac{1}{n}\sum_{i=1}^{n} \psi_j(x_i)$. Substituting back $\hat{a}_j$ in the orthonormal series expansion gives:

$$\hat{f}(x) = \frac{1}{n}\sum_{j=1}^{\infty}\sum_{i=1}^{n}\psi_j(x_i)\psi_j(x) \tag{2.12}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{\infty}\psi_j(x_i)\psi_j(x)\right) \tag{2.13}$$

$$= \frac{1}{n}\sum_{i=1}^{n}K(x, x_i). \tag{2.14}$$

Note that the term inside the brackets can be written as a kernel due to Mercer's theorem (52). Therefore, estimating a_j as the sample mean is equivalent to KDE if the basis results in a radial kernel (53).

The second method to estimate these coefficients adds the assumption that the true density is *sparse* in a given orthonormal basis. This assumption allows one to only consider finitely many basis functions as a support to the true density; and hence parametric ML estimation can be used for computing the coefficients. However, the ML solutions are hard to realize as the likelihood function is generally non-convex (in this setting) with multiple maxima. Therefore, optimization algorithms like gradient descent, Newton methods, or any other convex optimization procedures (54, 55) are used which at best guarantee convergence to local maxima and can take a long time to converge.

Finally, the third method maximizes the likelihood non-parametrically (infinite number of parameters) by choosing a proper basis function over which the non-parametric maximization can be done. The only known example to the author's knowledge for such maximization is the histogram (with rectangular basis functions).

Orthogonal Series Square-root Density Estimators Orthogonal series square-root density estimation (OSSDE) (56) is an extension of OSDE but it enforces positivity of the estimate explicitly, and hence is more parsimonious. Particularly, OSSDE assumes that the square-root of the unknown pdf is in the linear span of the orthonormal series, i.e.,

$$\sqrt{f}(x) = \sum_{j=1}^{\infty} a_j \psi_j(x). \tag{2.15}$$

Then, due to orthonormality of the basis functions, the coefficients a_j are

$$a_j =< \psi_j(x), \sqrt{f(x)} >= \mathrm{E}(\frac{\psi_j(x)}{\sqrt{f(x)}}). \tag{2.16}$$

Therefore, to estimate the a_j's, the pdf needs to be known a priori. Pinhiero et. al. (56) suggested to use a pre-estimator of $f(x)$, $\hat{f}(x)$, to estimate the coefficients. This process is non ideal and hence is seldom used in practice.

The second method to estimate the a_j's assumes that the pdf is sparse in the chosen orthogonal basis, and subsequently maximizes likelihood parametrically using numerical methods (57, 58). Peter et. al. (57) showed that the likelihood function's Hessian evaluated at the true parameter value converges to $(4\times)$ identity matrix as number of data points $n \to \infty$, and hence if n is large and one starts close to the true solution, the parametric ML method converges quickly. However, these two requirements are limiting and acceptable solutions can be found only for some choice of basis and underlying pdfs. In general, for these methods the likelihood function is non-convex and hence finding the ML solution

is tricky and time consuming and often results in sub-optimal solutions (local maximum) (58). Further, this method requires some prior knowledge about the choice of basis functions so that the sparsity assumption holds true. In cases where sparsity does not hold, these methods can lead to incorrect density estimates.

Finally, there is a possibility of implementing non-parametric ML under the square root formulation, however, doing this would need a proper choice of basis functions over which such a feat could be achieved. To author's knowledge, such an estimator is yet to be discovered and is the focus of Aim 1 of this thesis.

Wavelet Density Estimation and Sparsity With the discovery of wavelet transforms in 1990 (59), wavelets have become a natural choice for both OSDE and OSSDE basis functions (56, 60). For details about wavelets see (61). Wavelet basis functions have the advantage of setting up the smoothness at some level j_0, over which some local details (levels j_0 to j_1) can be added based upon the assumptions on the true pdf. Donho et. al. and Pinhiero et. al. (56, 60) suggested using a sparsity assumption to automatically add local details to the wavelet estimates for OSDE and OSSDE, respectively. In particular they suggested the use of soft (or hard) thresholds to scale (or set to zero) some of the wavelet coefficients. Donho et. al. also suggested that such threshold criteria minimize the maximum mean squared error over different functional spaces (e.g. Sobolev spaces). The proof for this is technical and not the scope of this thesis. The interested reader is directed to their original paper (60).

The relation of various non-parametric estimators to OSDE and OSSDE are shown in figures 2.1 and 2.2.

2.1.2.4 Penalized likelihood estimators

In the non-parametric case, it is seldom possible to maximize likelihood. For instance, consider the naive estimator that fits a dirac delta at each observed sample, i.e., $\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\delta(x - x_i)$. This results in a likelihood that is unbounded and hence cannot be maximized. Further, such estimates are useless from the point of view of generalizing on cross-validation data. To overcome this hurdle, penalized likelihood techniques (26) are used for non-parametric estimation. For these techniques, instead of maximizing the likelihood, a penalized likelihood function (penalized by a smoothness constraint) is often maximized. The cost functional, $J[f]$ for these methods can be written as:

$$J[f] = -\sum_{i}^{n} log(f(x_i)) + \lambda R[f]. \tag{2.17}$$

Here, the first term in the expression is standard negative log-likelihood whereas the second term penalize roughness of f. The parameter λ is used to set amount of acceptable roughness. A widely used roughness constraint is $R[f] = \int f''\mathrm{d}x$ (f'' is second derivative of f).

Although the penalized likelihood method yields a smooth estimate that also has high likelihood values, these methods do not produce ML estimators in the traditional sense and hence may not have nice asymptotic properties that ML methods have.

Finally, some approaches search over non-parametric sets for which an ML estimate exists. Some cases are discussed in (62, 63), wherein the authors construct maximum likelihood estimators for unknown but Lipschitz continuous pdfs. Although Lipschitz functions display desirable continuity properties, they can be non-differentiable. Therefore, such estimates can be non-smooth, but perhaps more importantly, they are not efficiently computable and a closed-form solution has not been derived yet (62, 63).

2.2 Estimation of dependencies between random variables

Dependencies quantify the maximum amount of information that can be extracted from one random variable about the other random variable. Estimating such dependencies is important in data science for selecting covariates that should be used for modeling the variable of interest. Ideal properties of a measure for dependency are described next.

2.2.1 Ideal properties of a dependencies measure

The ideal properties of a dependency measure have been in discussion since 1960s and concisely summarize by the Renyi's axioms (64, 65):

1. $\delta(X, Y)$ is defined for any pair of random variables X and Y, neither of them being constant with probability 1 (property of existence).

2. $\delta(X, Y) = \delta(Y, X)$ (property of symmetry).

3. $0 \leq \delta(X, Y) \leq 1$ (property of normalization of the second type).

4. $\delta(X,Y) = 0$ if and only if X and Y are independent (property of independence).

5. $\delta(X,Y) = 1$ if there is a strict dependence relationship between X and Y, i.e. either $X1 = \phi(Y)$ or $Y = \psi(X)$, where $\phi(\Delta)$ and $\psi(\Delta)$ are Borel-measurable functions.

6. If Borel-measurable functions $\phi(\Delta)$ and $\psi(\Delta)$ map the real axis in a one-to-one way onto itself, then $\delta(\phi(X,Y), \psi(X,Y)) = \delta(X,Y)$.

7. If the joint distribution of X and Y is normal, then $\delta(X,Y) = |\rho(X,Y)|$, where $\rho(X,Y)$ is the Bravais-Pearson linear correlation coefficient between X and Y.

This set of axioms was too restrictive for a dependency measure (65, 66) and amongst several dependency measures, the only measure which satisfies all axioms is the maximal correlation coefficient (64):

$$S(x,y) = \sup_{f,g}(\rho(f(X), g(Y))), \tag{2.18}$$

where the supremum is taken over all Borel functions f, g for which $r(f(X), g(y))$ is defined. Hall (67) later pointed out that S is not efficiently computable and takes the value of 1 too often. Schweizer (65) and later Granger (31) suggested modifications to Renyi's axioms that broaden them while keeping their essence. In particular, Axiom 6 is modified to $\delta(\psi(X), Y) = \delta(X,Y)$ for any strictly monotone continuous transformation ψ and Axiom 7 was modified to $\delta(X,Y) = f(|\rho|)$ for some simple function f, in the case of a bivariate Gaussian distribution. In addition, an extra property was added stating that the measure should ideally be a metric on densities. The list of Granger's axioms is given in Table 1.1.

2.2.1.1 Equitability and data processing inequality

In recent literature, the Granger Axiom 6 has mainly been replaced by a stronger notion called "self equitable" (68). Formally it states that:

Definition . A dependence measure Graphic is self-equitable if and only if it is symmetric (Graphic) and satisfies Formula

$$\delta(X, Y) = \delta(f(X), Y) \tag{2.19}$$

whenever f is a deterministic function, X and Y are variables of any type, and $X \leftrightarrow f(X) \leftrightarrow Y$ forms a Markov chain. Note f here is any deterministic function and not just a monotone function.

Any self equitable measure satisfies Granger Axiom 6. Further, Kinney et. al. also related self equitability with an even stronger but intuitive notion - the Data processing inequality (DPI) (69), which is stated below:

If $X \leftrightarrow f(X) \leftrightarrow Y$ forms a Markov chain, then a measure for dependency should obey the following equation:

$$\delta(X, Y) \geq \delta(X, Z) \tag{2.20}$$

Any DPI satisfying measure is self equitable and any F-Information measure (70) satisfies the DPI.

2.2.2 Popular measures of dependencies

In general, dependencies can be measured as some distance between the joint density and the product of marginals, i.e.,

$$\delta(X,Y) \triangleq ||f_{XY}(x,y) - f_X(x)f_Y(y)||. \tag{2.21}$$

It is straight forward to see that under this definition, Granger Axioms 1, 2, 7 are satisfied. Axiom 3 can be satisfied by simple normalization. However, measuring dependencies using the above mentioned definition require the knowledge of the joint and the marginal densities which are not readily available apriori. Therefore, simplifications of the above definition are proposed so as to estimate the dependencies directly from the data. Some popular simplifications (28, 29, 30, 32) are:

- Pearson's correlation

$$r \triangleq \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E(X)^2}\sqrt{E(Y^2) - E(Y)^2}}. \tag{2.22}$$

- Mutual information

$$I \triangleq \int \log\left(\frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}\right) f_{XY}(x,y)\mathrm{d}x\mathrm{d}y \tag{2.23}$$

- Distance correlation

$$\begin{aligned} \mathrm{dCov}^2(X,Y) &\triangleq \int \frac{|\phi_{XY}(s,t) - \phi_X(s)\phi_Y(t)|^2}{|s|^{1+p}|t|^{1+q}}\mathrm{d}s\mathrm{d}t \\ R &= \frac{\mathrm{dCov}(X,Y)}{\sqrt{\mathrm{dCov}(X,X)\mathrm{dCov}(Y,Y)}} \end{aligned} \tag{2.24}$$

here ϕ_{XY}, ϕ_X, ϕ_Y are the respective characteristic functions. p and q are the dimension of X and Y. For details see (29). (Note that, here the constants c_p, c_q are eliminated from the definition of dCov as they are not needed to define R).

- Mutual Dependence

$$d(X,Y) \triangleq d_h(f_{XY}(x,y), f_X(x)\, f_Y(y)) \tag{2.25}$$

with

$$d_h^2(p(\mathbf{x}), q(\mathbf{x})) \triangleq \frac{1}{2} \int \left(\sqrt{p(\mathbf{x})} - \sqrt{q(\mathbf{x})}\right)^2 \mathrm{d}\mathbf{x}. \tag{2.26}$$

Out of the four measures above the Pearson correlation is the simplest, but it is not a distance between joint density and the product of marginals, hence it only works well for linearly related random variables. Mutual Information is probably the most popular measure for measuring dependencies as it directly indicates the amount of information in bits that is contained in both X and Y. It is, however, not a strict distance metric but a divergence measure. Distance correlation, on the other hand, is a distance between the joint characteristic function and the product of marginals characteristic function and hence is very closely related to the notion of dependency defined previously. However, it has a priori fixed weights that penalize the mismatch at lower frequencies more and lvice-versa, which results in a biased measure of the dependencies. Finally, the mutual dependence is of desired form as it is defined as Bhattacharya Distance (71) between the joint and the product of marginals. Further it satisfies all the seven Granger's axiom and is a F-

information metric (see chapter 5) hence satisfy DPI and is self equitable. Therefore, in author's view, it is a very good candidate for measuring dependencies.

2.2.3 Estimation of theoretic measures

Apart from choosing the correct measure for measuring the dependencies a related and important issue is estimating the measure directly from the data. No matter how good a measure is but if it can not be estimated from data in a straight forward way it is of little use (e.g. maximal correlation coefficient). Both Pearson (28) and Distance correlations(29, 30) win over other measures in this category as there are theoretically nice ways to estimate them directly from data as shown below:

$$\hat{\rho} \triangleq \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.27}$$

here $\bar{x}$ and $\bar{y}$ are sample means i.e. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and

$$a_{jk} \triangleq ||x_j - x_k|| \tag{2.28}$$

$$b_{jk} \triangleq ||y_j - y_k|| \tag{2.29}$$

$$A_{jk} \triangleq a_{jk} - \frac{1}{n}\sum_{k=1}^{n} a_{jk} - \frac{1}{n}\sum_{j=1}^{n} \bar{a}_{jk} + \frac{1}{n^2}\sum_{j,k=1}^{n} a_{jk} \tag{2.30}$$

$$B_{jk} \triangleq b_{jk} - \frac{1}{n}\sum_{k=1}^{n} b_{jk} - \frac{1}{n}\sum_{j=1}^{n} \bar{b}_{jk} + \frac{1}{n^2}\sum_{j,k=1}^{n} b_{jk} \tag{2.31}$$

$$d\hat{C}ov(X,Y) \triangleq \frac{1}{n^2}\sum_{j,k=1}^{n} A_{jk}B_{jk} \tag{2.32}$$

$$\hat{R} \triangleq \frac{d\hat{C}ov(X,Y)}{\sqrt{d\hat{C}ov(X,X)d\hat{C}ov(Y,Y)}}. \tag{2.33}$$

For estimating other dependency measures one needs to estimate the joint and marginal densities first and then plug them in the formula of the corresponding measure and integrate. This process is error prone because first, the estimates of the densities are noisy and second, the integration have to be performed numerically which can amplify the estimation noise. Especially, in the case of mutual information the logarithm of the noisy density estimates have to be calculated which can amplify the noise tremendously (at low values of density). Mutual dependence is more stable in this regard as it only requires the calculation of square-root of the density. Further, the non-parametric estimators with square-root representation (like OSSDE) seems to be a natural choice for estimating the mutual dependence, however, such estimators have not previously been described in the literature.

The specific research aims of this thesis are described in next chapter.

Chapter 3

Research Aims

The research aims of this thesis are as follows.

AIM 1: To construct a non-parametric maximum likelihood smooth estimator for the square-root representation of a pdf. Under this aim a novel non-parametric ML estimator, that is both statistically and computationally efficiently computable, consistent and results in a smooth pdf is constructed. The pdf is assumed to have a finite-support in the Fourier domain, i.e., the pdf is band-limited (BL). The BL assumption can be thought of as a smoothness constraint. However, the proposed method will not require penalizing the likelihood function to guarantee the existence of a global maximum, and therefore may preserve the asymptotic properties of ML estimators (i.e. consistency, asymptotic normality and efficiency). The BLML estimator is then be applied to surrogate data generated from both BL and infinite-band pdfs and its performance is compared with KDE and OSSDE methods.

Aim 2: To construct a statistically and computationally efficient estimator

for mutual dependence using BLML estimator. In this aim the BLML estimator of mutual dependence, d, that computes d directly from the data is developed. The BLML estimator on substitution into the expression for mutual dependence (see (5.2)), results in a readily integrable closed form expression for mutual dependence. This avoids inaccurate and computationally inefficient numerical integration which is needed for estimating other dependency measures such as mutual information. Then another closed-form estimator for mutual dependence that uses OSSD estimator for density estimation is proposed and its performance with the BLML estimator for mutual dependence is compared. Finally, a thorough comparison (using data sets generated using different pdfs and types of linear and non-linear dependencies) of the BLML estimator for mutual dependence is done with the estimators for Pearson and distance correlation for accuracy, computational efficiency, and convergence rate.

Aim 3: To estimate the CIF for grid and place cells as a function of the rat's position and spike history covariates using BLML estimator . Further, decoding the trajectory of the rat in the circular arena by using spiking activity from grid and place cells and BLML estimator. In this aim, the BLML estimator is used to compute the CIF for hippocampal place cells and entorhinal cortex grid cells using Bayesian framework described previously. The performance of the BLML estimation is compared to KDE and GLM methods using test data. Note that this is very first attempt (to author's knowledge) to estimate the CIF for grid cells which has not been achieved previously in literature due to complex firing patterns of Grid cells. A novel deduction in

Bayesian framework is proposed that allowed successful estimation of CIF. This deduction further allowed the decoding of the trajectory of rat using the spiking activity of the place and grid cells, using both KDE and BLML methods. Both these methods are then compared on quality of the decoding performance.

Chapter 4

The BLML Estimator

In this chapter, a non-parametric maximum likelihood (ML) estimator for band-limited (BL) probability density functions (pdfs) is proposed. The consistency of BLML estimator is proved and approximate algorithms to compute it efficiently are developed. The BLML estimators are then compared with state-of-the-art non-parametric estimators (KDE and OSSDE) for both accuracy (integrated mean square error between estimated and true pdf) and computational speed.

4.1 The BLML estimator

The BLML estimator (72) is described in the following theorem.

Theorem 4.1.1 *Consider n independent samples of an unknown BL pdf, $f(x)$, with assumed cut-off frequency f_c. Then the BLML estimator of $f(x)$ is given as:*

$$\hat{f}(x) = \left(\frac{1}{n} \sum_{i=1}^{n} \hat{c}_i \frac{\sin(\pi f_c (x - x_i))}{\pi (x - x_i)} \right)^2, \tag{4.1}$$

where $\hat{\mathbf{c}} \triangleq [\hat{c}_1, \cdots, \hat{c}_n]^T$ *and*

$$\hat{\mathbf{c}} = \underset{\boldsymbol{\rho_n}(\mathbf{c})=\mathbf{0}}{\arg\max} \left(\prod_{i=1}^{n} \frac{1}{c_i^2} \right). \tag{4.2}$$

Here $\rho_{ni}(\mathbf{c}) \triangleq \frac{1}{n} \sum_{j=1}^{n} c_j s_{ij} - \frac{1}{c_i} \ \forall \ i = 1, \cdots, n$ *and*
$s_{ij} \triangleq \frac{\sin(\pi f_c(x_i - x_j))}{\pi(x_i - x_j)} \quad \forall \ i, j = 1, \cdots, n.$

Proof: See appendix A.

The system of equations, $\boldsymbol{\rho_n}(\mathbf{c}) = \mathbf{0}$ in (7.2) is monotonic, i.e., $\frac{\mathrm{d}\boldsymbol{\rho}_\mathbf{n}}{\mathrm{d}\mathbf{c}} > \mathbf{0}$, with discontinuities at each $c_i = 0$. Therefore, there are 2^n solutions, with each solution located in each orthant, identified by the orthant vector $\mathbf{c}_0 \triangleq \text{sign}(\mathbf{c})$. Each solution corresponds to a local maximum of the likelihood function which is also its maximum value in that orthant. Hence, the global maximum always exists and can be found by finding the maximum of these 2^n maxima. However, it is computationally exhaustive to solve (7.2), which entails finding the 2^n solutions of $\rho_\mathbf{n}(\mathbf{c}) = \mathbf{0}$ and then comparing values of $\prod \frac{1}{c_i^2}$ for each solution.

Therefore, efficient algorithms for the computation of the BLML estimator are developed.

4.1.1 Consistency of the BLML estimator

Proving consistency of the BLML estimator is not trivial as it requires a solution to (7.2). However, if $f(x) > 0 \ \ \forall x$ then consistency of BLML estimator can be established. To show this, first an asymptotic solution $\bar{\mathbf{c}}_\infty$ to $\boldsymbol{\rho_n}(\mathbf{c}) = \mathbf{0}$ is constructed (Theorem B.7.1). Then, consistency is established by plugging $\bar{\mathbf{c}}_\infty$ into (4.1) to show that the ISE and hence the

MISE between the resulting density, $f_\infty(x)$, and $f(x)$ is 0 (Theorem B.7.2). Then, it is shown that the KL-divergence between $f_\infty(x)$ and $f(x)$ is also 0, and hence $\bar{\mathbf{c}}_\infty$ is a solution to (7.2), which makes $f_\infty(x)$ the BLML estimator $\hat{f}(x)$ (Theorem B.7.3). These theorems and their proofs are presented in appendix B.

4.1.2 Generalization of the BLML estimators to joint pdfs

Consider the joint pdf $f(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^m$, such that its Fourier transform $F(\boldsymbol{\omega}) \triangleq \int f(\mathbf{x}) e^{-j\boldsymbol{\omega}^T \mathbf{x}} d\mathbf{x}$ has the element-wise cut off frequencies in vector $\boldsymbol{\omega}_c^{true} \triangleq 2\pi \mathbf{f}_c^{true}$. Then the BLML estimator is of the following form:

$$\hat{f}(x) = \left(\frac{1}{n} \sum_{i=1}^{n} \hat{c}_i \operatorname{sinc}_{\mathbf{f_c}}(\mathbf{x} - \mathbf{x}_i) \right)^2 \tag{4.3}$$

where, $\mathbf{f_c} \in \mathbb{R}^m$ is the assumed cutoff frequency, vector $\mathbf{x}_i' s \;\; i = 1 \cdots n$ are the data samples, $\operatorname{sinc}_{\mathbf{f_c}}(\mathbf{x}) \triangleq \prod_{j=1}^{m} \frac{\sin(\pi f_{cj} x_j)}{\pi x_j}$ and the vector $\hat{\mathbf{c}} \triangleq [\hat{c}_1, \cdots, \hat{c}_n]^T$, is given by

$$\hat{\mathbf{c}} = \underset{\boldsymbol{\rho_n}(\mathbf{c})=\mathbf{0}}{\arg\max} \left(\prod \frac{1}{c_i^2} \right). \tag{4.4}$$

Here $\rho_{ni}(\mathbf{c}) \triangleq \sum_{j=1}^{n} c_j s_{ij} - \frac{n}{c_i}$; $s_{ij} \triangleq \operatorname{sinc}_{\mathbf{f_c}}(\mathbf{x}_i - \mathbf{x}_j)$.

The multidimensional result can be derived in a very similar way as the one-dimensional result as described in appendix C.

4.1.3 Computing the BLML estimator

As discussed before, the BLML estimator is exponentially hard to compute in its raw form. Therefore three algorithms, `BLMLTrivial`, `BLMLQuick` and `BLML-BQP` are developed which are described next.

`BLML-BQP Algorithm.` This is a heuristic algorithm that first brings down the computational complexity of BLML estimator to that of a np-hard problem (73) and then again use heuristics to solve propose a polynomial time solution. To derive the `BLML-BQP` algorithm, it is first noted that the 2^n solutions of $\rho_n(\mathbf{c}) = \mathbf{0}$ are equivalent to the 2^n local solutions of:

$$\tilde{\mathbf{c}} = \underset{\mathbf{c}^T\mathbf{S}\mathbf{c}=n^2}{\text{arglocal max}}(\prod_i c_i^2). \tag{4.5}$$

here $\mathbf{S} \in \mathbb{R}^{n\times n}$ is a matrix with i,jth element being s_{ij}. Now, if $\mathbf{c}_0 \in \{1,-1\}^n$ is an orthant indicator vector and $\lambda \geq 0$ is such that $(\lambda\mathbf{c}_0)^T\mathbf{S}(\lambda\mathbf{c}_0) = n^2$, then (4.5) implies:

$$\prod_i \tilde{c}_i^2 \geq \lambda^{2n} \Rightarrow \prod_i \frac{1}{\tilde{c}_i^2} \leq \frac{(\mathbf{c}_0^T\mathbf{S}\mathbf{c_0})^n}{n^{2n}}. \tag{4.6}$$

Finally, the orthant where the solution of (7.2) lies is found by maximizing the upper bound $\frac{(\mathbf{c}_0^T\mathbf{S}\mathbf{c_0})^n}{n^{2n}}$ using the following binary quadratic program (BQP):

$$\hat{\mathbf{c}}_\mathbf{0} = \underset{\mathbf{c_0}\in\{-1,1\}^n}{\arg\max}\ (\mathbf{c}_0^T\mathbf{S}\mathbf{c_0}). \tag{4.7}$$

BQP problems are known to be NP-hard (74), and hence a heuristic algorithm implemented in the Gurobi toolbox (75) in MATLAB is used to find an approximate solution $\hat{\mathbf{c}}_0$ in polynomial time. Once a reasonable estimate for the orthant $\hat{\mathbf{c}}_0$ is obtained, $\rho_n(\mathbf{c}) = \mathbf{0}$ is solved in that orthant to find an estimate for $\hat{\mathbf{c}}$. To further improve the estimate, the solutions to $\rho_n(\mathbf{c}) = \mathbf{0}$ in all nearby orthants (Hamming distance equal to one) of the orthant $\hat{\mathbf{c}}_0$ are obtained and subsequently $\frac{1}{\tilde{c}_i^2}$ is evaluated in these orthants. The neighbouring orthant with the largest $\frac{1}{\tilde{c}_i^2}$ is set as $\hat{\mathbf{c}}_0$, and the process was repeated. This iterative process

is continued until $\frac{1}{\hat{c}_i^2}$ in all nearby orthants is no greater than that of the current orthant. The `BLML-BQP` is computationally expensive, with complexity $\mathcal{O}(n^2 + nl + BQP(n))$ where $BQP(n)$ is the computational complexity of solving BQP problem of size n. Hence, the `BLML-BQP` algorithm can only be used on data samples $n < 100$.

`BLMLTrivial Algorithm.` It is a one-step algorithm that first selects an orthant in which the global maximum may lie, and then solves $\rho_{\mathbf{n}}(\mathbf{c}) = \mathbf{0}$ in that orthant. As $\rho_{\mathbf{n}}(\mathbf{c}) = \mathbf{0}$ is monotonic, it is computationally efficient to solve in any given orthant.

As stated in Theorem B.7.4 (see appendix B), the asymptotic solution of (7.2) lies in the orthant with indicator vector $c_{0i} = 1 \; \forall i = 1, \cdots, n$ if $f(x)$ is BL and $f(x) > 0 \;\; \forall \; x \in \mathbb{R}$. Therefore, the `BLMLTrivial` algorithm selects the orthant vector $c_0 = \pm[1, 1, \cdots, 1]^T$, and then $\rho_{\mathbf{n}}(\mathbf{c}) = \mathbf{0}$ is solved in that orthant to compute $\hat{\mathbf{c}}$. It is important to note that when $f(x)$ is indeed BL and strictly positive, then the `BLMLTrivial` estimator converges to BLML estimator asymptotically.

Note that the conditions required by the BLMLTrivial algorithm are very less restrictive, as for sample sizes as few as 100 asymptotic effects can be observed, further the condition $f(x) > 0$ is obeyed by most pdfs encountered naturally. Therefore, BLMLTrivial or its derivative is the choice of algorithm to use in cases where no other information is available.

The computational complexity of the `BLMLTrivial` method is $\mathcal{O}(n^3 + nl)$ where l is the number of points where the value of pdf is estimated. This is very similar to computational complexity of KDE methods which is $\mathcal{O}(nl)$, (76)). As compared to KDE methods the

`BLMLTrivial` method has an extra step of solving equation $\rho_{\mathbf{n}}(\mathbf{c}) = \mathbf{0}$, which can be solved in n^3 computations using newton methods.

`BLMLQuick Algorithm.` The BL assumption of the true pdf allows for a quick implementation of the `BLMLTrivial` estimator - "`BLMLQuick`". For details, see appendix D. Briefly, `BLMLQuick` first groups the observed samples into bins of size $< \frac{0.5}{f_c}$. Then, it constructs the `BLMLTrivial` estimator of the discrete pdf (or the probability mass function, pmf) that generated the binned data. The true pmf for the binned data has infinite-bandwidth. Hence, under the required conditions, the `BLMLTrivial` estimate constructed using the Nyquist frequency, $2f_c$, converges to the continuous pdf $\bar{f}(x)$, from which the pmf is obtained via Nyquist-like sampling. $\bar{f}(x)$ can be made arbitrarily close to the true pdf $f(x)$ by choosing smaller and smaller bins. In fact, if the bin size reduces as $n^{-0.25}$, then the ISE between $\bar{f}(x)$ and $f(x)$ is of $\mathcal{O}(1/n)$. Therefore, the MISE for `BLMLQuick` is $\mathcal{O}(1/n)+$ MISE of the `BLMLTrivial` estimator. Since the MISE of the `BLMLTrivial` estimator has to be greater than $\mathcal{O}(1/n)$, the MISE for `BLMLQuick` algorithm is of same order as MISE for `BLMLTrivial` algorithm. The computational complexity of `BLMLQuick` is $\mathcal{O}(n + B^2 + lB)$, where $B \leq n$ is the number of bins. By considering, a $\frac{1}{x^r}$ tail for the true pdf the computational complexity becomes $\mathcal{O}\left(n + f_c^2 n^{0.5+2/(r-1)} + f_c n^{0.25+1/(r-1)} l\right)$. The derivation for the computational complexity is provided in appendix D.

4.2 Results

In this section, a comparison of `BLMLTrivial` and `BLML-BQP` algorithms on surrogate data generated from known pdfs is presented first. Then, the performance of the `BLMLTrivial` and `BLMLQuick` algorithms is compared to several KD estimators. Finally, the BLML estimator is compared with OSSD methods.

4.2.1 Performance of `BLMLTrivial` versus `BLML-BQP` on surrogate data

In figure 4.1, `BLMLTrivial` and `BLML-BQP` estimates are presented assuming that the true pdfs are BL by $f_c = f_c^{true}$. Panels (A, C) and (B, D) use surrogate data generated from a non-strictly positive pdf $f_x = 0.4\,\mathrm{sinc}^2(0.4x)$ and strictly positive pdf $f(x) = \frac{3\times 0.2}{4}(\mathrm{sinc}^4(0.2x) + \mathrm{sinc}^4(0.2x + 0.1))$, respectively. Both pdfs are BL from $(-0.4, 0.4)$. In Panels A and B, the BLML estimates ($n = 81$) are plotted using both algorithms, and the true pdfs are overlaid for comparison. In Panels C and D, the MISE is plotted as a function of sample size n for both algorithms and both pdfs. For each n, data were generated 100 times to generate 100 estimates from each algorithm. The mean of the ISE was then taken over these 100 estimates to generate the MISE plots.

As expected from theory, the `BLML-BQP` algorithm works best for the non-strictly positive pdf, whereas the `BLMLTrivial` algorithm is marginally better for the strictly positive pdf. Note that as n increases beyond 100, the `BLML-BQP` algorithm becomes computationally expensive, therefore the `BLMLTrivial` and `BLMLQuick` algorithms are used in the remainder of this paper with the assumption that the true pdf is strictly positive.

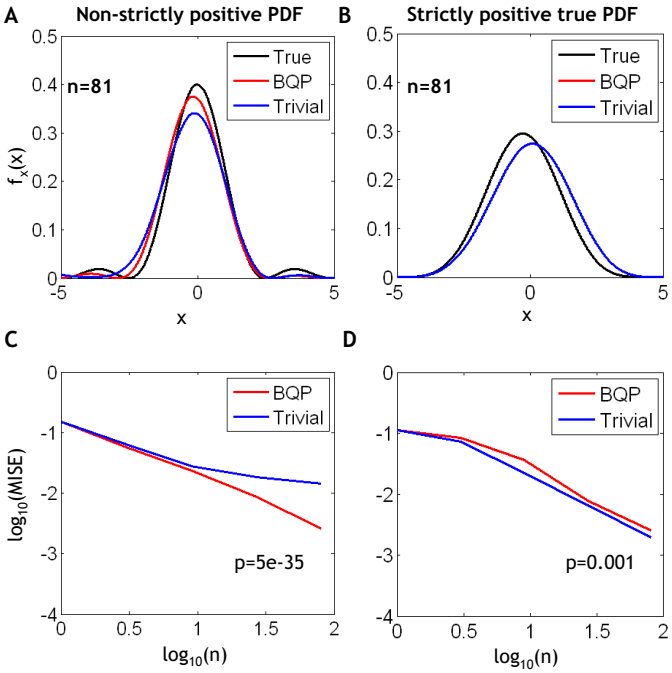

Figure 4.1: Comparison of `BLMLTrivial` and `BLML-BQP` - Illustration of the results of `BLMLTrivial` and `BLML-BQP` algorithms using a non-strictly positive true pdf $f(x) = 0.4\,\mathrm{sinc}^2(0.4x)$, (**A,C**) and a strictly positive pdf $f(x) = \frac{3\times 0.2}{4}(\mathrm{sinc}^4(0.2x) + \mathrm{sinc}^4(0.2x + 0.1))$, (**B,D**). The cut-off frequency was assumed to be $f_c = f_c^{true}$. The p-values were calculated using a paired t-test at $n = 81$. Note in **(B)**, the red line is beneath the blue line.

4.2.2 BLML and KDE on surrogate data

The performance of the `BLMLTrivial` and `BLMLQuick` estimates is compared with adaptive KD estimators which are the fastest known non-parametric estimators with convergence rates of $\mathcal{O}(n^{-4/5})$, $\mathcal{O}(n^{-12/13})$ and $\mathcal{O}(n^{-1})$ for 2nd-order Gaussian (KDE2nd), 6th-order Gaussian (KDE6th) and sinc (KDEsinc) kernels, respectively (48, 77). Panels A and B of figure 4.2 plot the MISE of the BLML estimators using the `BLMLTrivial`, `BLMLQuick`, and

the adaptive KD approaches for cases in the presence of BL or non-BL pdf, respectively. In the BL case, the true pdf is strictly positive and is the same as used above, and for the infinite-band case, the true pdf is *normal*. For the `BLMLTrivial`, `BLMLQuick` and sinc KD estimates, $f_c = 2f_c^{true}$ and $f_c = 2$ are used for the BL and infinite-band cases, respectively. For the 2nd and 6th-order KD estimates, the optimal bandwidths ($q = \frac{0.4}{f_c} n^{-1/5}$ and $q = \frac{0.4}{f_c} n^{-1/13}$ respectively) are used. The constant $\frac{0.4}{f_c}$ ensures that MISEs are matched for $n = 1$.

It can be seen from the figure that for both the BL and infinite-band cases, `BLMLTrivial` and `BLMLQuick` outperform KD methods. In addition, the BLML estimators seem to achieve a convergence rate that is as fast as the KDEsinc, which is known to have a convergence rate of $\mathcal{O}(n^{-1})$. Figure 4.2 C plots the MISE as function of the cut-off frequency f_c for the BL pdf. `BLMLTrivial` and `BLMLQuick` seem to be most sensitive to the correct knowledge of f_c, as it shows larger errors when $f_c < f_c^{true}$, which quickly dip as f_c approaches f_c^{true}. When $f_c > f_c^{true}$, the MISE increases linearly and the BLML methods have smaller MISE as compared to KD methods.

Finally, figure 4.2D plots the computational time of the BLML and KD estimators. All algorithms were implemented in MATLAB, and in-built MATLAB 2013a algorithms were used to compute the 2nd and 6th-order adaptive Gaussian KD and sinc KD estimators. The results concur with theory and illustrate that `BLMLTrivial` is slower than KD approaches for large number of observations, however, the `BLMLQuick` algorithm is remarkably quicker than all KD approaches and `BLMLTrivial` for both small and large n.

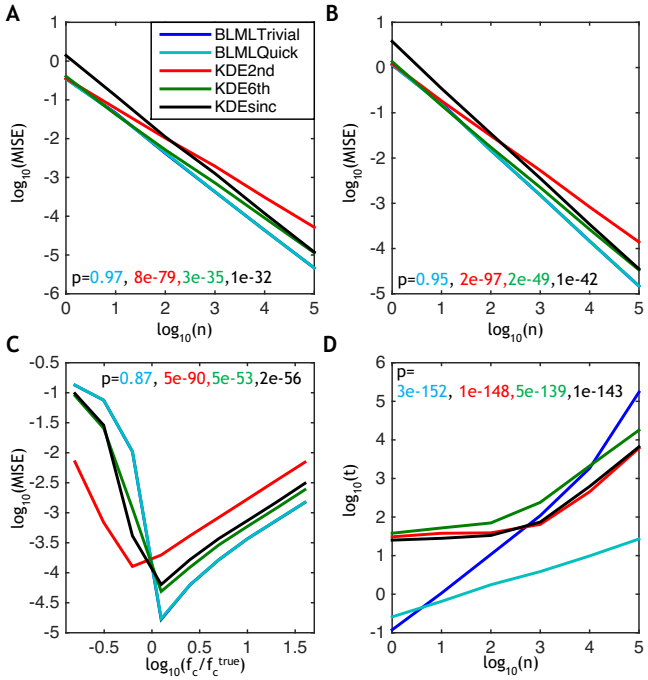

Figure 4.2: Comparison of BLML and KD estimation - Comparision of the results of the `BLMLTrivial` and `BLMLQuick` estimators to the KDE2nd, KDE6th and sinc KD estimators. MISE as a function of n for (**A**) a strictly positive band-limited true pdf (the one used in figure 4.1 B) and (**B**) *an infinite band standard normal pdf*. For the BLML estimators the cut-off frequencies are chosen as $f_c = 2f_c^{true}$ for the BL true pdf and $f_c = 2$ for the normal true pdf. For the KDE2nd and KDE6th, the optimal bandwidths were chosen as $q = \frac{0.4}{f_c} n^{-0.2}$ and $\frac{0.8}{f_c} n^{-1/13}$, respectively and also to match the MISE for the BLML estimator for $n = 1$. For the KDEsinc, the f_c was kept the same as the f_c for BLML estimators. (**C**) MISE as a function of the cut-off frequency $\frac{f_c}{f_c^{true}}$ for a BL true pdf with cut-off frequency f_c^{true}. $n = 10^4$ was used for creating this plot. (**D**) Computation time as a function of n. The p-values were calculated between the `BLMLTrivial` estimator and other estimators using paired t-tests for either $\log_{10}(n) = 5$ (**A,B,D**) or $\log_{10}(f_c/f_c^{true}) = 1.6$ (**C**) and are color coded.

4.3 BLML and OSSDE on surrogate data

In this section the performance of the BLMLQuick and the OSSD estimators is compared. For fair comparison Fourier series basis is chosen for OSSDE estimator. In particular the following form is used:

$$\sqrt{f}(x) = a_0 + \sum_{i=1}^{M} a_i cos(2\pi f_0 i) + \sum_{i=1}^{M} b_i sin(2\pi f_0 i) \tag{4.8}$$

here $\frac{1}{f_0}$ is the assumed support (period) of the density in time domain, a_i, b_is are the unknown coefficients. $2M + 1$ are total number of coefficients. Sparsity, in particular band-limitedness then allowed to assume a finite M (In this setting the cut-off frequency $f_c = 2Mf_0$) and parametric maximum likelihood is then performed as described in (57).

For comparison with BLML methods the surrogate data is generated from standard normal density with zero mean and unit variance. Therefore, it is assumed safely that density has support in $(-5, 5)$ in time domain, and hence f_0 was set to 0.1 whenever required. The simulation results are presented in figure 4.3. Panel 4.3A plots MISE for the BLMLQuick and OSSDE methods as a function of number of samples n. For generating the plot $f_c = 2$, 2 $f_0 = 0.1$ and $M = 10$ is assumed. The panel clearly show that BLML methods has significantly lower MISE than OSSDE methods for same number of data points. The convergence rates are also better for BLML methods, however OSSDE seems to achieve a similar convergence rate as the number of samples grow. This may be due to the fact that the the likelihood function under OSSDE formulation has hessian matrix (calculated at the parameter values that reflect the true density) approaches (4x) identity matrix as number

of sample points increases (57). Panel 4.3B plots MISE as a function of assumed cut-off frequency f_c and $M = \frac{f_c}{2f_0}$. $f_0 = 0.1$ for reasons described previously. It can be seen from the panel that MISE was intially lower for OSSDE method but as assumed cut-off frequency is increased MISE became lower for BLML method which shows a minimum MISE at around $f_c = 0.5$. This value corresponds well with the power spectrum of the standard normal density which have almost all power inside $f_c = 0.5$. The reason that OSSDE performed better in the beginning is probably because at very low cut-off frequencies M was equal to 1 making the OSSDE class very restrictive. The normal densities are close to this restrictive class and hence maximizing likelihood yielded a better MISE, than BLML methods. Panel 4.3C and 4.3D plots MISE and computational time as a function of f_0, as BLML methods does not depend on choice of f_0, MISE is a constant line for BLML methods. For OSSDE methods the MISE does not depend on f_0 either as the normal density can be safely thought as having support in $(-5, 5)$ and hence all $f_0 > 0.1$ are equivalent. More importantly, it can be seen that MISE for BLML methods is much lower that OSSDE methods, this reflects the fact the OSSDE methods do not guarantee a ML solution as the optimization algorithm can easily produce a local maximum. Finally, reducing f_0 to very low values results in increase of M, (for a given cut-off frequency) which results in the increase in computational time as shown in panel 4.3D.

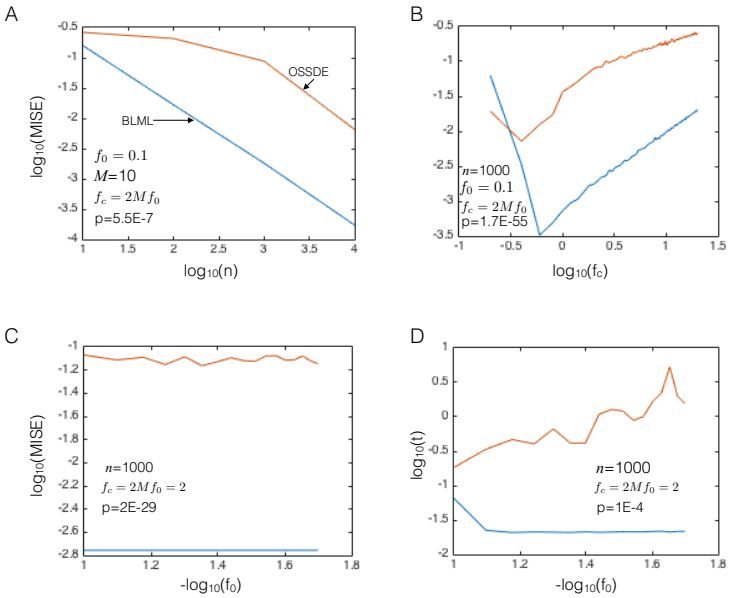

Figure 4.3: Comparison of BLML and OSSD estimation - In this plot the simulation results of the `BLMLQuick` and OSSD estimators are compared. Comparison of the MISE as a function of (**A**) n. for $f_c = 2$ (**B**) assumed cut-off frequency f_c for $n = 1000$. In both the plots $f_0 = 0.1$ was used for the reasons described in text and M was set so as to make $2Mf_0 = f_c$. (**C**) The MISE as a function of f_0. $f_c = 2$, $N = 1000$ are used for creating this panel. (**D**) Computation time as a function of f_0 for a fixed $f_c = 2$. Therefore M scaled as $\frac{f_c}{f_0}$. The p-values were calculated between the `BLMLQuick` estimator and OSSD estimators using paired t-tests for $\log_{10}(n) = 4$, (**A**), $f_c = 20$ (**B**) and $f_0 = 0.02$ (**C,D**).

4.4 Discussion

In this chapter, a non-parametric ML estimator for BL densities is developed and its consistency is proved. In addition, three heuristic algorithms that allow for quick computation of the BLML estimator are presented. Although these algorithms are not guaranteed to generate the BLML estimate, it is shown that for strictly positive pdfs, the

`BLMLTrivial` and `BLMLQuick` estimates converge to the BLML estimate asymptotically. Further, `BLMLQuick` is significantly quicker than all tested KD methods, while maintaining convergence rates of BLML estimators. Even further, using surrogate data, it is shown that both the `BLMLTrivial` and `BLMLQuick` estimators have an apparent convergence rate of $1/n$ for MISE, which is equal to that of parametric methods.

The BLML estimators may also be motivated by quantum mechanics. The function $g(x)$ in the development of BLML estimate (see appendix A) is analogous to the wave function (78) in quantum mechanics, where the square of the absolute value of both are probability density functions. In addition, in quantum mechanics the wave function of momentum is the Fourier transform of the wave function of position. Therefore, if the momentum wave function has finite support, then the position wave function is BL and vice versa. Such occurrences are frequent in the single or double slit experiment, where one observes band-limited ($\text{sinc}^2(f_1 x)$ and $\cos^2(f_2 x)$ respectively) profile for the probability of finding a particle at a distance x from the center. Also, in the thought experiment of a particle in a box: the wave function for position has finite support, making the momentum wave function BL. Author suspect that a large number pdfs in the nature are BL because macro world phenomenon are a sum of quantum level phenomenon and pdfs at quantum level are shown to be BL (single and double slit experiments). Furthermore, the set of BL pdfs is complete, i.e. the sum of two random variables that each have a BL pdf is a random variable whose pdf is a convolution of original pdfs, and hence is BL. Therefore, if macro level phenomenon is a linear combination of different quantum level phenomenon with BL

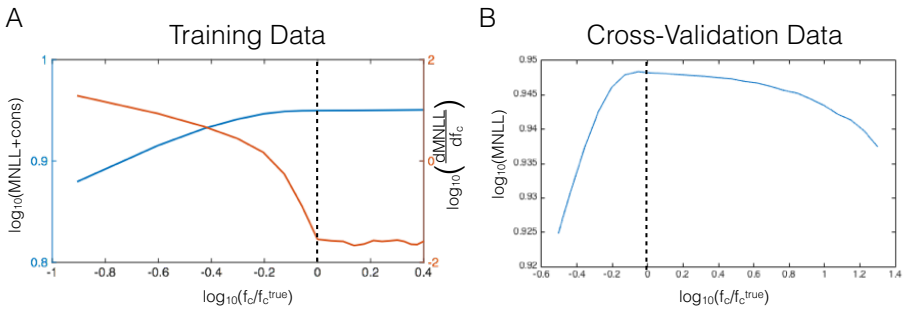

Figure 4.4: Estimation of f_c^{true} - MNLL and $\frac{dMNLL}{df_c}$ curves as a function of f_c. The cons is an arbitrary constant that is added to MNLL so that the logarithm of sum could exist.

pdfs, then the macro level phenomenon will also generate a BL pdf. In fact, we see this at macro level where we observe Gaussian pdfs of various processes. The Gaussian pdf is almost BL, with cut-off frequency $f_c = 10/\sigma$ ($< 10^{-324}\%$ of its power lies outside this band). In fact, given finite data, it is impossible to distinguish if the data is generated by a Gaussian or BL pdf.

4.4.1 Choosing a cut-off frequency for the BLML estimator

The BLML method requires selecting a cut-off frequency of the unknown pdf. One strategy for estimating the true cut-off frequency is to first fit a Gaussian pdf using the data via ML estimation. Once an estimate for standard deviation is obtained, one can estimate the cut-off frequency using the formula $f_c = 1/\sigma$, as this will allow most power of the true pdf to lie within the assumed band if the true pdf has Gaussian-like tails.

Another strategy is to increase the assumed cut-off frequency of BLML estimator as a function of the sample size. For such a strategy, the BLML estimator may converge even

when the true pdf has an infinite frequency band, provided that the increase in cut-off frequency is slow enough and the cut-off frequency approaches infinity asymptotically, e.g. $\omega_c \propto \log(n)$.

A more sophisticated strategy would be to look at the mean normalized log-likelihood (MNLL), $\mathbb{E}(-\frac{1}{n}\sum \log(\hat{c}_i^2))$ as a function of assumed cut-off frequency f_c. Figure 4.4A plots MNLL (calculated using `BLMLTrivial` algorithm) is plotted for $n = 200$ samples from a strictly positive true pdf $f(x) = \frac{3\times 0.2}{4}(\text{sinc}^4(0.2x) + \text{sinc}^4(0.2x + 0.1))$ along with $\frac{d\text{MNLL}}{df_c}$. Note that $\frac{d\text{MNLL}}{df_c} \simeq \mathbb{E}(\frac{1}{n^2}\sum_{ij} \hat{c}_i\hat{c}_j o_{ij})$, where $o_{ij} \triangleq \cos(f_c(x_i - x_j))$. It can be seen that the MNLL rapidly increases until f_c reaches f_c^{true}, after which the rate of increase sharply declines. There is a clear "knee" in both MNLL and $\frac{d\text{MNLL}}{df_c}$ curves at $f_c = f_c^{true}$. Therefore, f_c^{true} can be inferred from such a plot. A more complete mathematical analysis of this "knee" is left for future work.

Finally, one can use cross-validation procedure for selecting the cut-off frequency. In particular one can calculate the normalized likelihood values as a function of assumed cut-off frequency using cross-validation data set as shown in figure 4.4B. As shown in the figure, such normalized likelihood attain a maximum or near maximum value near true cut-off frequency which can be used to infer the true cut-off frequency. Further, the plot shows that the mean normalized likelihood value decays quite slowly if the true cut-off frequency is over estimated. This in turn show that the BLML methods are very robust to the choice of assumed cut-off frequency as long as it is greater than the true cut-off frequency.

4.4.2 Making `BLMLQuick` even faster

There are several faster implementation of KD approaches such as those presented in (76, 79). These approaches use numerical techniques to evaluate the sum of n kernels over l given points. Such techniques may also be incorporated while calculating the `BLMLQuick` estimator to make it even faster. Exploration of this idea will be done in a future study.

4.4.3 Asymptotic properties of the BLML estimator

Although, this paper proves that the BLML estimate is consistent, it is not clear whether it is asymptotically normal and efficient (i.e., achieving a Cramer-Rao-like bound). Studying asymptotic normality and efficiency is non-trivial for BLML estimators as one would need to first redefine asymptotic normality and extend the concepts of Fisher information and the Cramer-Rao lower bound to the non-parametric case. Therefore, this is left for a future study. However, the author postulate here that the curvature of MNLL plot might be related to Fisher information in the BLML case. In addition, although under simulations, the BLML estimator seems to achieve a convergence rate similar to its parametric counterparts ($\mathcal{O}_p(n^{-1})$) its theoretical proof is also left for a future study.

Chapter 5

Mutual Dependence and its Estimator

In data science, it is often required to estimate dependencies between different data sources. These dependencies are typically calculated using Pearson's correlation, distance correlation, and/or mutual information. However, none of these measures satisfy *all* of Granger's axioms for an "ideal measure". One such ideal metric, proposed by Granger himself, calculates the Bhattacharyya distance between the joint pdf and the product of the marginal pdfs. This metric is called the *mutual dependence* in this thesis. However, to date this metric has not been shown to be directly computable from data and required first estimating the joint and marginal densities from data samples, and then numerical integration. Because of mathematical form of Bhattacharya distance the error prone and computationally expensive numerical integration might not be required if density estimators that use square-root representation are used. Therefore the OSSD and BLML estimators seem to be the

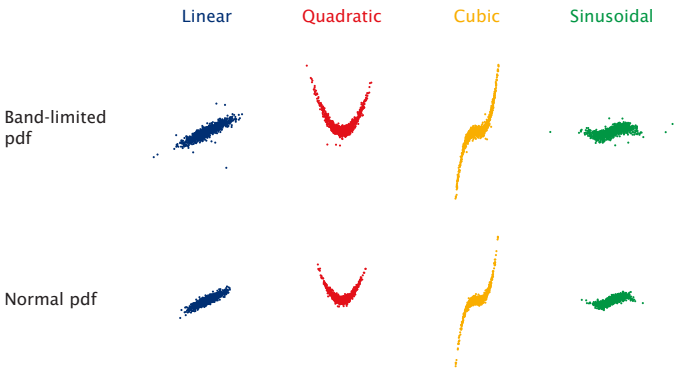

Figure 5.1: Point clouds - Illustrating point cloud for data generated from (5.1) for different nonlinearities $g(x)$ and generating pdfs. $\rho = 0.9$ was used for generating this data.

natural choice for computing the mutual dependence. In this chapter, both the OSSD and BLML estimator are used to derive an explicit representation for the mutual dependence as a function of the data. Then, the performance between these two estimators is compared via simulations. Finally, the convergence rates and computational complexity of the BLML estimator of mutual dependence is compared to standard measures (Pearson's and distance correlation) for different non-linear dependencies and generating pdfs.

5.1 Motivating example

Consider two random variables X and Y defined as:

$$X = V,$$
$$Y = \rho\, g(X) + \sqrt{1-\rho^2}\, U,$$

where U and V are two random variables that can have any unknown distributions. For

the purpose of this example lets consider that both U and V follow the same distribution that can either be the following band-limited pdf

$$f_X(x) = \frac{3}{4}\left[\text{sinc}^4(0.2\,x - 0.1) + \text{sinc}^4(0.2\,x + 0.1)\right]$$

or the following normal pdf

$$f_X(x) = \frac{1}{\sigma\sqrt{2\,\pi}} e^{-\frac{(x-\mu)^2}{2\,\sigma^2}}$$

and where $g(X)$ is one of the following four functions

$$X,\ X^2,\ X^3, \text{ or } \sin(X).$$

The 'spread' ρ is varied from 0.1 to 0.9 to obtain different degrees of dependencies. Figure 5.1 illustrates data generated for this example.

The amount of dependencies between X and Y can be accurately captured by mutual information as it obeys the Granger's axioms 1, 2, 5, and 6, which are probably the most important of his 7 axioms (the other axioms deal with normalization and scaling). Further, in case of surrogate data where underlying pdfs are known the mutual information is also readily computable. Therefore, for now, the mutual information is used as a bench mark to compare other dependency measures with. Figure 5.2, plots theoretical values for Pearson's and distance correlation of dependence as a function of mutual information for the four different non-linearity types and the two different generating pdfs. Figure shows that both Pearson's and distance correlation show significantly different values (depending on the type of non-linearity) for dependencies that have similar mutual Information value.

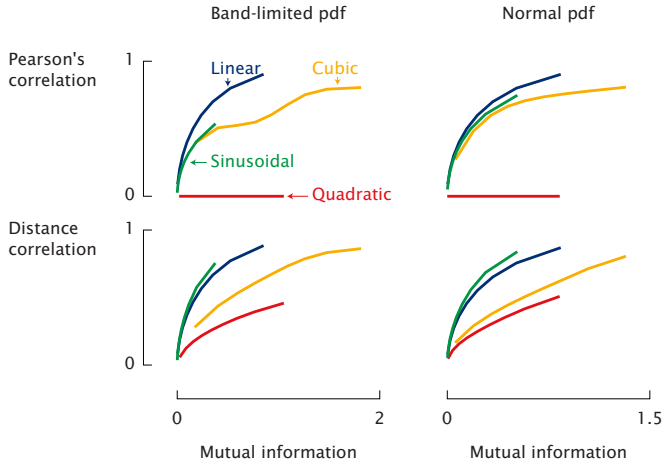

Figure 5.2: Pearson's and distance correlation - Illustrating theoretical values of r and R as a function of I for different nonlinearities and generating pdfs as used in figure 5.1.

This variability may occur because both correlation measures are not invariant to strictly monotonic transformations (Granger's Axiom 6) and/or obey the data processing inequality (DPI), unlike mutual information. Therefore, changing the type of non-linearity results in different values for both Pearson's and distance correlations, while the mutual information remains invariant. Such variance is undesirable as it may lead to incorrect inferences when comparing dependencies between data having different types of non-linear dependencies. Therefore, a measure that is invariant to strictly monotonic transformations is desirable.

5.2 Mutual dependence and its estimation

In this section, *mutual dependence* is introduced and its properties are stated. Then, two estimators for mutual dependence are derived that rely on BLML and OSSD density esti-

mators. Finally, efficient algorithms to compute these estimators are described.

5.2.1 Mutual dependence

Consider two random variables X and Y, their joint distribution $f_{XY}(x, y)$, and their marginal distributions $f_X(x)$ and $f_Y(y)$. These random variables are independent if and only if $f_{XY}(x, y) = f_X(x)\, f_Y(y)$. It is therefore natural to measure dependence as the distance (in the space of pdfs) between the joint and the product of marginal distributions. A good distance candidate is the Bhattacharyya distance (also known as Hellinger distance). See (31, 80) for details.

Definition 1 *The* mutual dependence $d(X, Y)$ *between two random variables X and Y is defined as the Bhattacharyya distance $d_h(\cdot, \cdot)$ between their joint distribution $f_{XY}(x, y)$ and the product of their marginal distributions $f_X(x)$ and $f_Y(y)$, that is,*

$$d(X, Y) \triangleq d_h(f_{XY}(x, y), f_X(x)\, f_Y(y)) \tag{5.2}$$

with

$$d_h^2(p(\mathbf{x}), q(\mathbf{x})) \triangleq \frac{1}{2} \int \left(\sqrt{p(\mathbf{x})} - \sqrt{q(\mathbf{x})} \right)^2 \mathrm{d}\mathbf{x}. \tag{5.3}$$

In this thesis, this measure is called "mutual dependence" as it represents mutual information most closely. Further the mutual dependence is also tightly related with mutual information as for a given value of mutual information, the value of mutual dependence remains almost the same irrespective of the non-linearity type see figure 5.3.

5.2.2 Properties of mutual dependence

Due to symmetry of $d(\cdot, \cdot)$, it is easy to see that $d(X, Y) = d(Y, X)$. The measure $d \in (0, 1)$ if X and Y are partially dependent which quantifies the degree of dependence between the

two random variables. In the extreme cases, $d = 0 \iff X$ and Y are independent and $d = 1$ if either x or y is a Borel-measurable function of the other. Also, it can be easily established that d is invariant under strictly monotonic transformations ψ_1 and ψ_2, i.e., $d(X, Y) = d(\psi_1(X), \psi_2(Y))$. A detailed description of these properties can be found in (31, 80).

For jointly normal data, the mutual dependence can be estimated by first calculating the Bhattacharyya distance between two multivariate Gaussian distributions (81)

$$d_h^2 = 1 - \frac{|\Sigma_1|^{\frac{1}{4}} |\Sigma_2|^{\frac{1}{4}}}{\left|\frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2\right|^{\frac{1}{2}}} \times \exp\left(-\frac{1}{8}(\mu_1 - \mu_2)^T \left(\frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2\right)^{-1} (\mu_1 - \mu_2)\right) \tag{5.4}$$

where μ_1 and μ_2 are the mean vectors and Σ_1 and Σ_2 covariance matrices. Then substituting

$$\mu_1 = 0, \qquad \Sigma_1 = \begin{bmatrix} \sigma_x^2 & \rho\,\sigma_x\,\sigma_y \\ \rho\,\sigma_x\,\sigma_y & \sigma_y^2 \end{bmatrix},$$

$$\mu_2 = 0, \qquad \Sigma_2 = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix},$$

gives

$$d(X, Y) = \sqrt{1 - \frac{(1-\rho^2)^{\frac{1}{4}}}{(1 - \frac{1}{4}\rho^2)^{\frac{1}{2}}}} \triangleq M(\rho). \tag{5.5}$$

This shows that mutual dependence satisfies axiom 5 (see Table 1).

Finally, mutual dependence is also an F-information measure (70), and therefore obeys the DPI and is also self equitable (68). Briefly, a F-information measure can be written as:

$$\delta(X, Y) = \int F\left(\frac{f_{xy}(x, y)}{f_x(x) f_y(y)}\right) f_x(x) f_y(y) \mathrm{d}x\mathrm{d}y \tag{5.6}$$

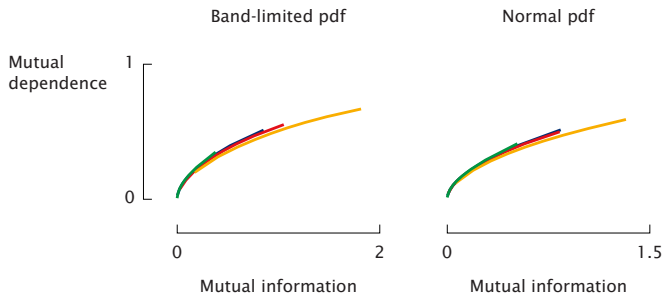

Figure 5.3: Mutual dependence - Illustrating theoretical values of d as a function of I for different nonlinearities and generating pdfs as used in figure 5.1.

where f_{xy}, f_x, f_y are the joint and marginal pdfs and $F()$ is a convex function on non-negative real numbers. It is straight forward to see that mutual dependence is also an F-information measure with function $F(x) = (1 - \sqrt{x})$.

5.2.3 Estimation of mutual dependence

To estimate d, the OSSDE and BLML methods (72) described in chapters 2 and 3 are used, respectively. The reason for using OSSDE and BLML estimation methods is due to the fact that the structure of these estimators is well suited for evaluating the integral in (5.2), resulting in an estimate which is a direct function of observed data and hence avoids numerical integration errors.

Consider OSSDE. The OSSD estimator for mutual dependence can directly be obtained by substituting the the OSSD density estimator into the expression for the mutual dependence. This result is described in the following theorem.

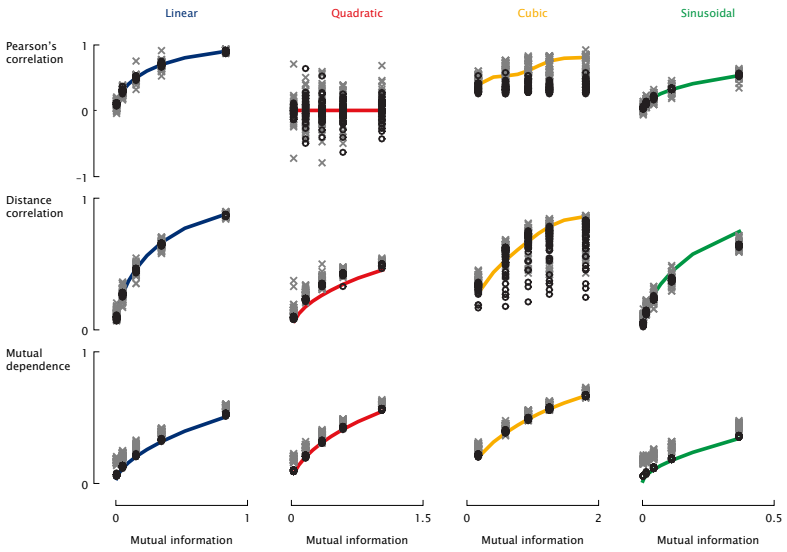

Figure 5.4: Monte Carlo Estimates for band-limited generating pdfs - The Monte Carlo distribution of estimates for different measures for different non-linearities and band-limited generating pdfs. ×ess mark the estimates calculated using sample sizes $n = 316$ whereas ○s mark the estimates calculated using sample size $n = 10000$. d is estimate assuming the cut-off frequency $f_c = \frac{1}{1-\rho^2}$.

Theorem 5.2.1 *Let* (x_i, y_i) $i = 1, \cdots, n$ *be* n *paired independent and identically distributed data observations. Then the OSSD estimator for mutual dependence is given as:*

$$\hat{d}_{OSSD} \triangleq d_h(\hat{f}_{XY}, \hat{f}_X \hat{f}_Y) = \sqrt{1 - \sum_{i=1}^{\infty} \sum_{i=1}^{\infty} \hat{a}_{ij}^{(XY)} \hat{a}_i^{(X)} \hat{a}_i^{(Y)}} \tag{5.7}$$

where $\hat{a}_{ij}^{(XY)} \hat{a}_i^{(X)}, \hat{a}_j^{(Y)}$ *are the the corresponding coefficients of the OSSD estimator for joint and marginal densities.*

See (72) for details.

The BLML estimator can also be used to estimate d as shown in the following theorem (82).

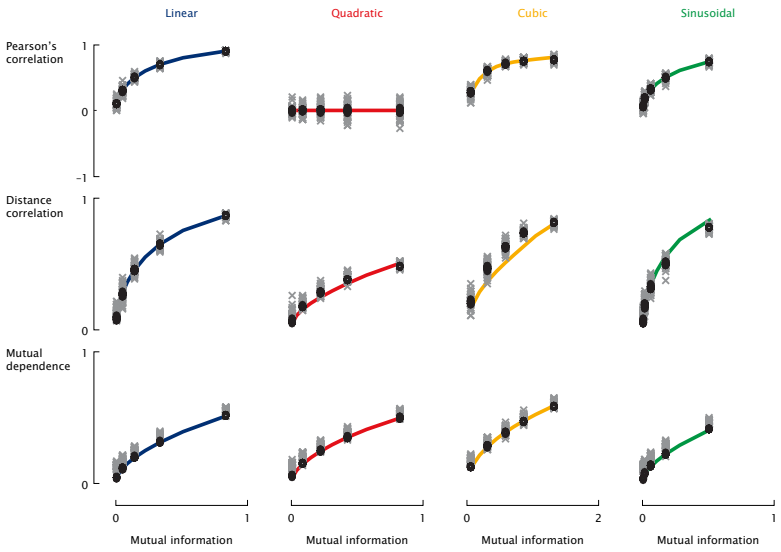

Figure 5.5: Monte Carlo Estimates for normal generating pdfs - The Monte Carlo distribution of estimates for different dependency measures for different non-linearities and the normal generating pdf. ×es mark the estimates calculated using sample sizes $n = 316$ whereas ○s mark the estimates calculated using sample size $n = 10000$. d is estimate assuming the cut-off frequency $f_c = \frac{1}{1-\rho^2}$.

Theorem 5.2.2 *If* $(x_i, y_i)\ \ i = 1, \cdots, n$ *are* n *paired i.i.d data observations and* f_c *is the cut-off frequency parameter. Then the BLML estimator for mutual dependence is given as:*

$$\hat{d} \triangleq d_h(\hat{f}_{XY}, \hat{f}_X \hat{f}_Y) = \sqrt{1 - \frac{1}{n} \sum \frac{\hat{c}_i^{(XY)}}{\hat{c}_i^{(X)} \hat{c}_i^{(Y)}}} \tag{5.8}$$

where $\hat{\mathbf{c}}^{(XY)} = \{\hat{c}_i^{(XY)}\}_{i=1}^n$ *is given by:*

$$\hat{\mathbf{c}}^{(XY)} = \underset{\boldsymbol{\rho}_{\boldsymbol{n}}^{(\boldsymbol{XY})}(\mathbf{c})=\mathbf{0}}{\arg\max} \left(\prod \frac{1}{c_i^2} \right)$$

$$\rho_{ni}^{(XY)}(\mathbf{c}) \triangleq \sum_{j=1}^{n} c_j \frac{\sin(\pi f_c(x_i - x_j))}{\pi(x_i - x_j)} \frac{\sin(\pi f_c(y_i - y_j))}{\pi(y_i - y_j)} - \frac{n}{c_i},$$

$\hat{\mathbf{c}}^{(X)} = \{\hat{c}_i^{(X)}\}_{i=1}^n$ *is:*

$$\hat{\mathbf{c}}^{(X)} = \underset{\boldsymbol{\rho}_{\boldsymbol{n}}^{(\boldsymbol{X})}(\mathbf{c})=\mathbf{0}}{\arg\max} \left(\prod \frac{1}{c_i^2}\right)$$
$$\rho_{ni}^{(X)}(\mathbf{c}) \triangleq \sum_{j=1}^{n} c_j \frac{\sin(\pi f_c(x_i - x_j)}{\pi(x_i - x_j)} - \frac{n}{c_i}$$

and $\hat{\mathbf{c}}^{(Y)} = \{\hat{c}_i^{(Y)}\}_{i=1}^n$ *is:*

$$\hat{\mathbf{c}}^{(Y)} = \underset{\boldsymbol{\rho}_{\boldsymbol{n}}^{(\boldsymbol{Y})}(\mathbf{c})=\mathbf{0}}{\arg\max} \left(\prod \frac{1}{c_i^2}\right)$$
$$\rho_{ni}^{(Y)}(\mathbf{c}) \triangleq \sum_{j=1}^{n} c_j \frac{\sin(\pi f_c(y_i - y_j))}{\pi(y_i - y_j)} - \frac{n}{c_i}.$$

Proof The BLML estimators of f_{XY}, f_X and f_Y from Theorem 4.1.1 (using same cut-off frequency $[f_c, f_c]$, f_c and f_c respectively) are plugged into (5.2) and the resultant equation is integrated which gives $\hat{d}$.

In chapter 4, the BLML estimator is shown to outperform OSSDE and other non-parametric methods, such as kernel density estimators, both in convergence rates and computational time and hence provides a better alternative for non-parametric estimation of dependency measures. Further, it gives a closed form expression for $\hat{d}$ and is of direct use for the estimation of dependencies between data.

5.2.4 Consistency of $\hat{d}_{BLML}$

Consistency of $\hat{d}_{BLML}$ is straightforward to prove. In (72) we have shown that if $f > 0$ and $f_c > f_c^{true}$, $||\hat{f} - f||_2 \overset{a.s}{\to} 0$, which implies $d_h(\hat{f}, f) = 0$. Therefore,

$$\begin{aligned}
\lim_{n\to\infty} \hat{d}_{BLML} &\triangleq \lim_{n\to\infty} d_h(\hat{f}_{xy}, \hat{f}_x\hat{f}_y) \\
&= \lim_{n\to\infty} \left\| \sqrt{\hat{f}_{xy}} - \sqrt{\hat{f}_x\hat{f}_y} \right\|_2 \\
&\leq \lim_{n\to\infty} \left(\left\| \sqrt{\hat{f}_{xy}} - \sqrt{f_{xy}} \right\|_2 + \left\| \sqrt{f_{xy}} - \sqrt{\hat{f}_x\hat{f}_y} \right\|_2 \right) \\
&\leq \left\| \sqrt{f_{xy}} - \sqrt{f_x f_y} \right\|_2 + \lim_{n\to\infty} \left\| \sqrt{f_x f_y} - \sqrt{\hat{f}_x\hat{f}_y} \right\|_2 \overset{a.s.}{\to} d
\end{aligned}$$

Similarly, it can be shown that $d_{BLML} \overset{a.s.}{\leq} \lim_{n\to\infty} \hat{d}$, hence $d \overset{a.s.}{\to} \lim_{n\to\infty} \hat{d}$. Therefore: $\hat{d}_{BLML} = 1 - \int \sqrt{\hat{f}_{xy}(x,y)\hat{f}_x(x)\hat{f}_y(y)}\mathrm{d}x\mathrm{d}y \to 1 - \int \sqrt{f_{xy}(x,y)f_x(x)f_y(y)}\mathrm{d}x\mathrm{d}y = d.$ ∎

5.2.5 Computation of BLML estimator for mutual dependence

As described in (72) solving for $\hat{\mathbf{c}}$ requires exponential time. Therefore, heuristic algorithms also described in (72) such as `BLMLBQP` and `BLMLTrivial`, can be used directly to compute $c_i^{(XY)}$, $c_i^{(X)}$, $c_i^{(Y)}$ approximately for each i for small scale ($n < 100$) and large scale ($n > 100$) problems, respectively.

To further improve computational time, `BLMLQuick` algorithm (72) can also be used. `BLMLQuick` uses binning and estimates $c_i^{(XY)}$, $c_i^{(X)}$, $c_i^{(Y)}$ approximately for each i. It is also shown in (72) that both `BLMLTrivial` and `BLMLQuick` algorithms yield consistent estimate of pdfs if the true pdf is strictly positive, therefore in cases where the joint $f_{XY} > 0$, the estimate, d is also consistent.

5.2.6 Estimation of the cut-off frequency

As suggested in (72), when estimating univariate pdfs, a good choice for cut-off frequency is $\frac{\gamma}{\sigma}$ (σ is the standard deviation of data, γ is bandwidth of standard normal density; good values that can be used are 0.4 or 1), for bivariate pdfs this corresponds to $\frac{\gamma}{\sigma_{min}}$, where σ^2_{min} is the least singular value (83) of the covariance matrix Σ^2.

$$\Sigma^2 = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

Here, σ_x, σ_y are the standard deviations in x and y directions respectively and ρ is the correlation coefficient. The covariance matrix is able to describe data well when dependency is linear however in cases when dependency is non-linear covariance matrix is not a good metric, for such cases covariance matrix can be generalized yielding generalized covariance matrix Σ_g^2 as follows:

$$\Sigma_g^2 = \begin{pmatrix} \sigma_x^2 & \rho_g\sigma_x\sigma_y \\ \rho_g\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

here ρ_g is generalized correlation coefficient (GCC), which is defined as:

$$\rho_g \triangleq \sqrt{\frac{\sqrt{64 - 48(1-d^2)^4} + 4(1-d^2)^4 - 8}{(1-d^2)^4}} \tag{5.12}$$

which reduces to ρ for bivariate Gaussian data. Therefore, a reasonable choice of cut-off frequency is $\frac{\gamma}{\sigma_{min}}$, where σ^2_{min} is the least singular value of the generalized covariance matrix Σ_g^2. However, for estimating Σ_g^2 one needs to know d, therefore the following algorithm is proposed:

1. Initialize $f_c = \frac{\gamma}{\sigma_{min}}$ of Σ and $dp = 0$.

2. Estimate d using one of the BLML algorithms.

3. Find Σ_g and its least singular value σ_{min} (square root of least eigenvalue of Σ_g^2).

4. If $d - dp > 0$. Update $f_c = \frac{1}{\sigma_{min}}$ and $dp = d$ go to step 2.

5. Output $\hat{d}_{BLML} = d$. Stop.

This algorithm is called GCC algorithm from here onwards in this thesis.

5.3 Performance of the estimators for mutual dependence

In this section, the performance of OSSD and BLML estimators for mutual information is first compared. The performance of the BLML estimator for mutual dependence is compared with estimators for Pearson's and distance correlation by comparing the empirical distribution of the estimator with the empirical distribution of the estimators for Pearson's and distance correlation for different mutual information values, I, non-linearities, $g(X)$, and generating pdfs, $f_x(x)$. Then the mean squared errors between the of the BLML estimator for mutual dependence, the estimator for Pearson and distance correlations and their true values are compared for different sample sizes. Finally, the comparison of computational complexity of our estimator with the estimator for distance correlation is done.

5.3.1 Comparison of the BLML and OSSD estimators for mutual dependence

For a fair comparison between BLML and OSSD estimators the Fourier series basis for OSSD is chosen and maximum likelihood estimation is done for estimating the coefficients. This method is described in detail in previous chapter in section 4.3. The same approach is used for estimating the joint density.This yields:

$$\begin{aligned}\sqrt{f}_{xy}(x,y) = &\sum_{i=0}^{M}\sum_{j=0}^{M} a_{ij}^{(XY)} \cos(2\pi f_0 xi)\cos(2\pi f_0 yj) + \sum_{i=0}^{M}\sum_{j=1}^{M} a_{ij}^{(XY)} \cos(2\pi f_0 xi)\sin(2\pi f_0 yj) \\ &+ \sum_{i=1}^{M}\sum_{j=0}^{M} a_{ij}^{(XY)} \sin(2\pi f_0 xi)\cos(2\pi f_0 yj + \sum_{i=1}^{M}\sum_{j=1}^{M} a_{ij}^{(XY)} \sin(2\pi f_0 xi)\sin(2\pi f_0 yj)\end{aligned}$$

As can be seen the number of coefficients becomes $(2M+1)^2$ for 2-d densities as opposed to $2M+1$ for $1-d$ densities. Therefore, OSSD methods even after using computationally fast algorithms (suggested by Peter et. al. (57)) have trouble in converging for the 2-d densities and Fourier series representation. For data generated from true joint normal density with zero correlation, $f_0 = 0.05$, $M = 4$, $f_c = 0.4$ resulted in an OSSDE algorithm that converged most of the times, and also these parameters nicely represent the true density as most of the power of jointly normal distribution lie inside the assumed band $|f_c| < 0.4$. Therefore these parameter values are chosen for the comparison with BLML estimators for mutual dependence. The BLML estimators only need the cut-off frequency parameter which is set to $f_c = 0.4$, to match that of OSSD estimators.

Figure 5.6 plots the joint (panels 5.6A, 5.6C) and product of marginal densities (panels 5.6B, 5.6D) computed by the OSSD estimators (panels 5.6A, 5.6B) and BLML estimators

(5.6C, 5.6D) for data generated ($n = 1000$) from jointly normal density with zero correlation. For such data the product of marginals and the joint density are exactly equal to each other making theoretical $d = 0$. However, the OSSD estimators for the joint and product of marginals show estimation bias and do not directly corresponds to each other (panels 5.6A and 5.6C). This may be because of the reasons that the OSSD estimators do not directly depend on the data points as the BLML or KDE methods do and there is no guarantee that the OSSD methods converge to the global maximum of likelihood. This mismatch is a big factor for the estimation of mutual dependence as the mutual dependence essentially measures the distance between product of marginals and the joint density which can become bias due to such mismatch. On the other hand the BLML methods do not suffer from such estimation bias and one can see from panel 5.6C, 5.6D that the estimators for the joint and product of marginals are very similar to each other resulting a quite accurate estimator of d. For these reasons the $\hat{d}_{BLML}$ is used for estimating the mutual dependence for here onwards and is denoted by $\hat{d}$.

5.3.2 Comparison of convergence rate for different nonlinearities

Figures 5.4 and 5.5 plot the estimated $\hat{r}$, $\hat{R}$ and $\hat{d}$ for $n = 316$ and $n = 10000$ from about 50 Monte Carlo runs as a function of I for different non-linearities (linear, quadratic, cubic and sinusoidal) and generating pdfs (band-limited and normal). Underlaid are the respective theoretical values. Specifically, the first row shows about 50 Monte Carlo computation of $\hat{r}$ for different I values, non-linearities and generating pdfs. It can be seen that for both

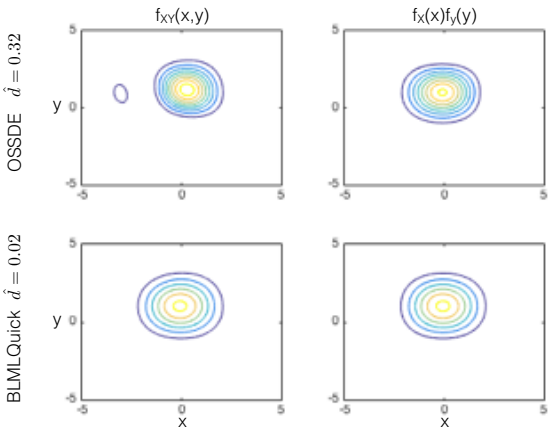

Figure 5.6: Comparison between BLML and OSSD estimator for mutual dependence - **(A,C)**The estimated joint desnity and **(B,D)** the product of marginals using the **(A,B)** OSSD and **(C,D)** BLML methods. The estimated $\hat{d}$ values are mentioned on the left. The data $n = 1000$ was generated from jointly normal density with zero correlation.

$n = 316$ and $n = 10000$, $\hat{r}$ works best for linear and sinusoidal data, but for quadratic data $\hat{r}$ has a larger variance and for cubic data $\hat{r}$ has a larger bias in bandlimited case. The second row shows 50 Monte Carlo computations of $\hat{R}$ for different I values, non-linearities and generating pdfs. It can be seen that for both $n = 316$ and $n = 10000$, $\hat{R}$ works best for linear data, but for quadratic and sinusoidal data, it has larger bias whereas for cubic data it has larger variance. The bottom row shows 50 Monte Carlo computation of $\hat{d}$ for different I values, non-linearities and generating pdfs. It can be seen that $\hat{d}$ works equally good for all non-linearities and shows less bias and variance than both $\hat{r}$ and $\hat{R}$.

Figure 5.7 plots the integration (over different I values) of mean squared error (IMSE) between the theoretical and estimated measures using about 50 Monte Carlo runs, for

different non-linearities and generating pdf types.

$$IMSE = \int \frac{1}{m} \sum_{i=1}^{m} (\hat{\delta}(I) - \delta(I))^2 \, \mathrm{d}I \tag{5.13}$$

Here, m is the number of Monte Carlo simulations and δ is the dependency metric. It can be seen from the figure 5.7 that the convergence rate is fastest for $\hat{d}$ irrespective of non-linearity type and/or generating pdf. r and R show an equally fast convergence rate for linear and normal data, but the rate is slower for non-linear and non-normal data. Specifically, the first row shows convergence of $\hat{r}$, from which it can be established that convergence of $\hat{r}$ to the theoretical values is fastest for linear data. For non-linear data, the convergence is slow either due to large bias or variance as discussed previously. The second row shows convergence of $\hat{R}$. It can be seen that $\hat{R}$ does well for linear data, but the rate slows down and saturates for non-linear data again due to either large bias or variance. Specially, for cubic and band-limited data, the IMSE of $\hat{R}$ does not decrease with increasing the number of samples, this is due to the non-decreasing variance of the estimator (see figure 5.4). The bottom row shows convergence of $\hat{d}$. It can be seen that $\hat{d}$ converges equally well for all data types and generating pdfs.

5.3.3 Comparison of computational complexity

The computational complexity of computing $\hat{r}$ is least which is $\mathcal{O}(n)$, whereas computational complexity of computing $\hat{R}$ is maximum which is $\mathcal{O}(n^2)$. $\hat{d}$ is same as computational complexity of `BLMLQuick` algorithm which is $\mathcal{O}(B^2 + n)$, where B is the number of bins

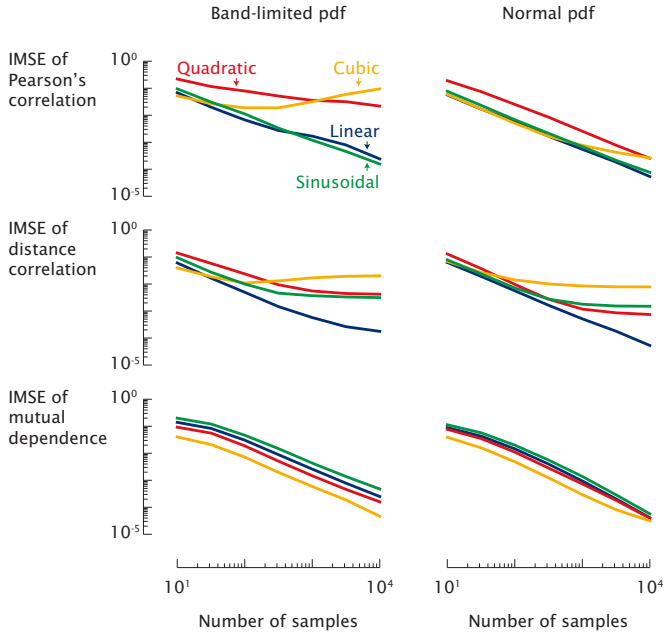

Figure 5.7: Integrated mean squared error vs sample size - Showing the Integrated mean squared error as a function of sample size n for different measures, different non-linearities and different generating pdfs. d is estimate assuming the cut-off frequency $f_c = \frac{1}{1-\rho^2}$.

containing non-zero number of samples, which is always less than equal to n. For dense data $B \ll n$ therefore computation of $\hat{d}$ is a lot quicker than estimating $\hat{R}$ in such cases.

5.4 Dependence of the BLML estimator for mutual dependence on assumed cut-off frequency

Figure 5.8 plots the d values as a function of assumed cut-off frequency for different n values for data generated from band-limited true joint pdf and having linear dependence. A knee can be observed at the point of true cut-off frequency, which becomes sharper and sharper

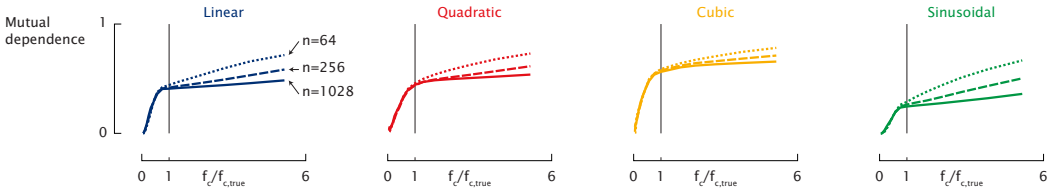

Figure 5.8: Dependence of $\hat{d}$ on the cut-off frequency - Plotting $\hat{d}$ as a function of assumed cut-off frequency for data generated from band-limited true joint pdf and four different non-linearities $g(x) = x, x^2, x^3, sin(x)$ **(A,B,C,D)**. It can be seen that in all four panels a knee is observed at around the true cut-off frequency of the joint pdf.

as n is increased.

5.4.1 Performance of GCC algorithms for estimating the true cut-off frequency

Figure 5.9 plots the convergence results of the GCC algorithm for data generated from different generating pdfs with different non-linear dependencies. Specifically, figure 5.9 A plots the d as a function of f_c for data generated from band-limited and normal true pdf with quadratic and cubic dependencies respectively. Overlaid on top of these plots are the (d, f_c) points over which the GCC algorithm iterates. Figure 5.9 B ,C shows the evolution of d and f_c respectively, for the data generated from band-limited and normal pdf for the four types of non-linearities $(x, x^2, x^3, sin(x))$. It can be seen that irrespective of the generating pdf and the non-linearity the GCC algorithm provide reasonable convergence results to the $\hat{d}$ values that are close to true d values.

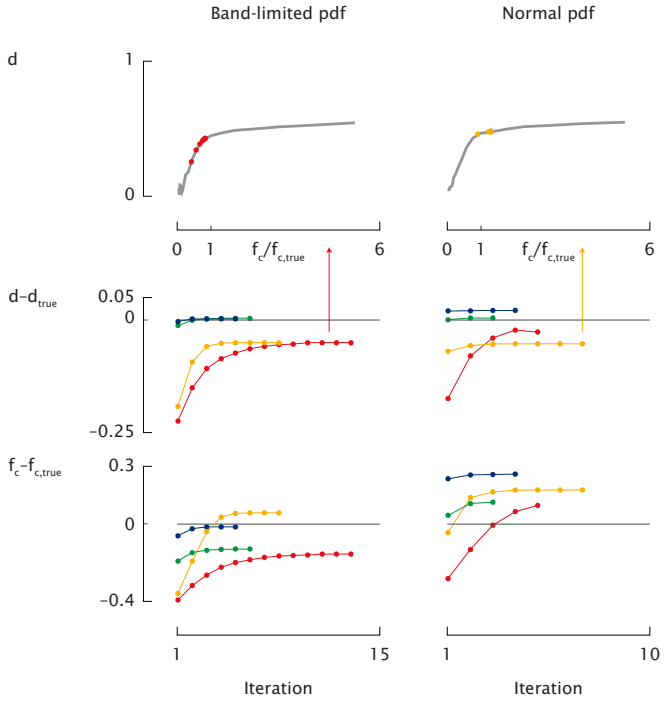

Figure 5.9: Convergence results for GCC algorithm for determining the cut-off frequency - **(A)** d as a function of f_c for data generated from band-limited and normal true pdf with quadratic and cubic dependencies respectively. Over laid on top of this are the dots indicating the iteration points (d, f_c) of GCC algorithm. The **(B)** d and **(C)** f_c values taken by GCC algorithm as a function of iteration number for the data generated from band-limited and normal pdf and four different non-linearities.

5.4.2 Application to Wikipedia data

Figure 5.10 plots the r, R, and d metrics calculated for the different data types posted on the Wikipedia page of correlation and distance correlation. The first row contains the linearly dependent data generated from jointly Gaussian pdf, the second row contains totally dependent data with different rotations, third and fourth row contains different type of non

linear dependencies with low and high amount of dependencies respectively. It can be seen from the figure that Pearson's correlation fails to capture non-linear dependence, which is captured successfully by distance correlation and mutual dependence. Further, on a careful look the distance correlation also seem to have self contradictory values as panels in last row 1st, 5th and 6th column has lesser estimated distance correlation than panel in first row 2nd and 3rd column. The mutual dependence, on the other hand does not seem to contradict itself across the range of different data used in the Wikipedia figure.

5.5 Conclusions

In this chapter, a novel estimator for measuring the mutual dependence is introduced. This estimator uses the BLML estimator and computes the mutual dependence directly computed from the data. The mutual dependence is bhattacharya distance between the joint and marginal densities and is an "ideal" measure for dependence between two random variables (31). Further, it is proved the mutual dependence is a F-information measure and hence satisfy data processing inequality and is self-equitable. These properties are not satisfied by state-of-the-art measures for dependencies such as Pearson correlation, distance correlation and mutual information and hence the estimator introduced in this chapter has clear advantages over the state-of-the-art estimators of dependencies. In particular, the BLML estimator for the mutual dependence has advantages over mutual information estimators as mutual dependence is a metric, has values in $[0, 1]$. It also has advantages over Pearson's and distance correlation estimators as first the mutual dependence is invariant

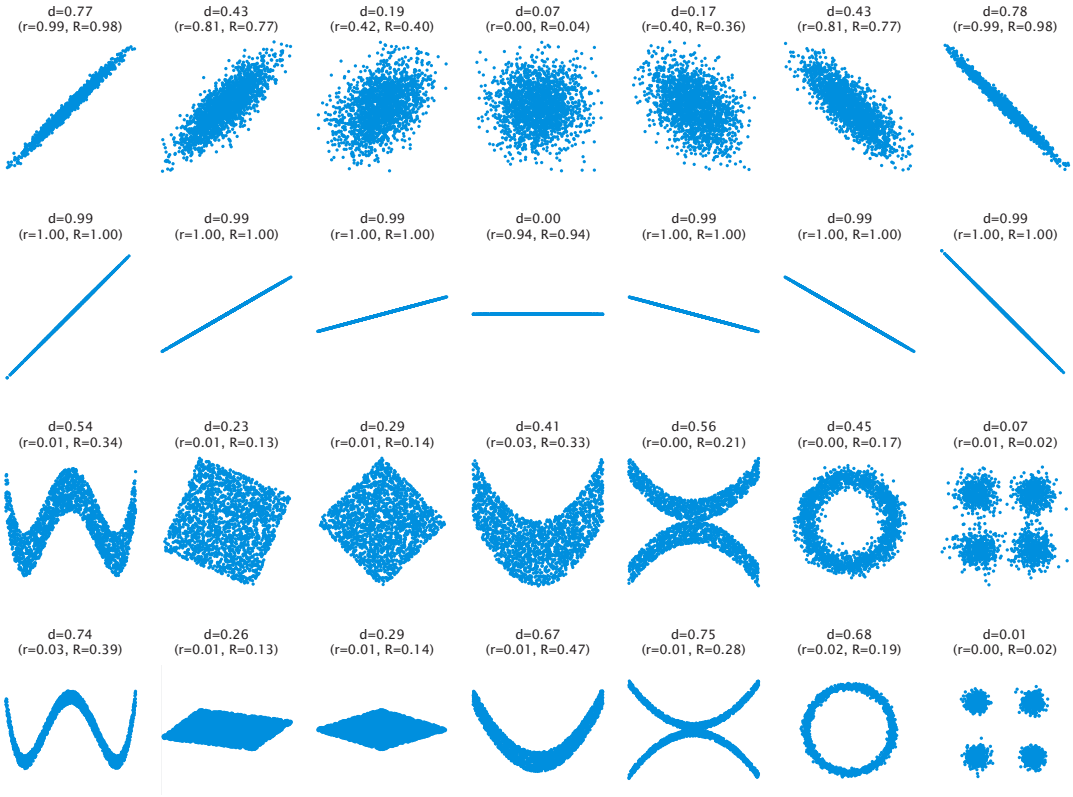

Figure 5.10: Performance of $\hat{d}$, $\hat{r}$, $\hat{R}$ on data from Wikipedia - Data from Wikipedia page on Pearson and distance correlation pages is used to compare the performance of the BLML estimator of mutual dependence with the estimators for Pearson and distance correlations. The GCC algorithm is used to select the cut-off frequency for the BLML estimator for mutual dependence.

under invertible transformation. Second, under simulation, estimators of both Pearson's and distance correlation require more samples to achieve the same integrated mean squared error (IMSE) as compared to the BLML estimator for mutual dependence for a variety of non-linear dependencies and generating pdfs. The slower convergence rate for the estimators

of Pearson's and distance correlation was due to their higher variance and bias for the non-linearly dependent data. Such non-linearities did not affect the BLML estimator for mutual information and it showed a uniform decrease in IMSE as the sample size increases for all tested non-linearities. Third, the BLML estimator for mutual dependence showed a computational time complexity of $\mathcal{O}(B^2 + n)$ where $B \leq n$ is the number of bins, which is superior to the time complexity of distance correlation ($\mathcal{O}(n^2)$) and is much faster when the data is dense.

Chapter 6

Application of the BLML Estimator to Grid and Place Cells

As mentioned in the introduction the "Fourier hypothesis" (22) hypothesize that the place cell fields are formed by the Fourier like summation of grid cells periodic fields. This make the BLML estimator a natural candidtae for estimating the encoding fields of place and grid cells. Therefore, in this chapter the "Fourier hypotheis" is tested. For this, BLML, KD, and GLM methods are applied to estimate the CIF for place and grid cells from spike train data and the better two models (chosen through a model section criterion) are used to decode the rat's trajectory from spike trains generated by these cells. If the Fourier hypothesis is valid than it is anticipated that the BLML methods will perform atleast as good as other modeling techniques, both in encoding and decoding.

6.1 Methods

6.1.1 Data collection

In the experimental set up, the Long-Evans rat was freely foraging in an open field arena of radius of 1m for a period of 30-60 minutes. Custom micro-electrode drives with variable numbers of tetrodes were implanted in the rat's medial entorhinal cortex and dorsal hippocampal CA1 area. Spikes were acquired with a sampling rate of 31.25 kHz and filter settings of 300 Hz-6 kHz. Two infra-red diodes alternating at 60 Hz were attached to the micro-electrode array drive of the animal for position tracking. Spike sorting was accomplished using a custom manual clustering program (Xclust, M.A. Wilson). All procedures were approved by the MIT Institutional Animal Care and Use Committee.

In total, 74 neurons were recorded from both entorhinal cortex and hippocampus. Out of these 74 neurons, 21 neurons were discarded as they fire either independently of position x, y (inter-neurons) or they have very low firing rates $< 0.175Hz$. Out of remaining 53 neurons, 27 were classified as unimodal place cells and 26 were classified as multi-modal place cells or grid cells. The x and y coordinates of rat's position when the uni-modal place cells, multi-modal place cells, and grid cells spiked are shown as scatter plots in figures 6.1 and 6.2.

6.1.2 Model estimation

As discussed in the chapter 1, the spiking activity of a place or grid cell (and any neuron in general) is modeled as a point process whose pdf on any random variable defined on

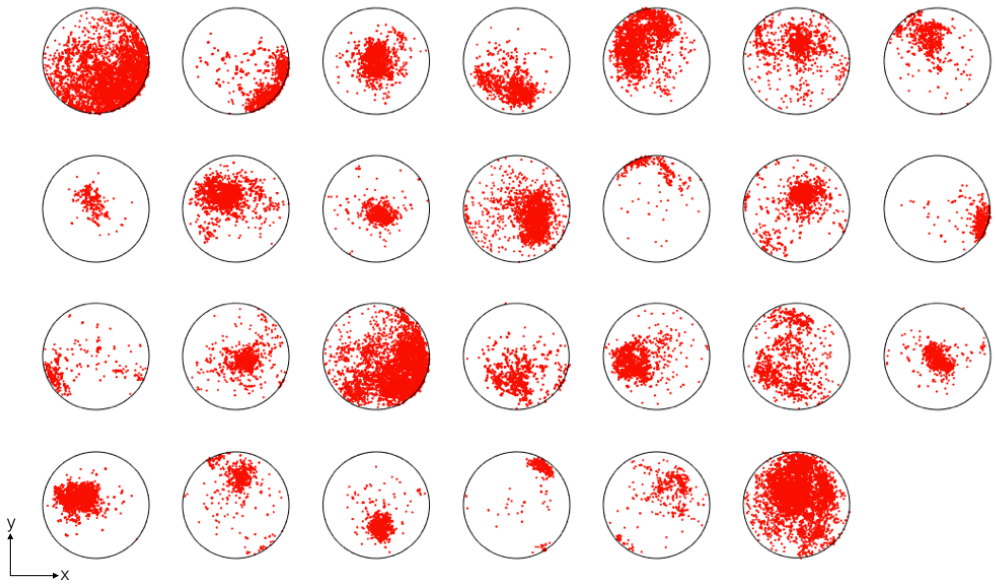

Figure 6.1: Spike scatter plot for uni-modal place cells - Each plot corresponds to spiking activity of a single place cell displayed in x, y co-ordinates in the circular arena. Each dot corresponds to x, y coordinates of rat's position when the cell spiked.

the process (e.g. inter spike interval) is entirely characterized by the conditional intensity function (CIF). The spiking activity of place and grid cells are known to be modulated by the rat's position (4, 5), whose peak firing locations define a place field or a grid-like array covering much of the 2-dimensional arena. Further, the spiking activity of place and grid cells might also depend on spiking history due to the relationship of hippocampus and entorhinal cortex to memory formation (2, 84, 85, 86, 87). In particular Fenton et. al. showed (2) that the place cell firing pattern variability is higher so as to be explained only by the dependence on the location of the rat. Therefore, in this thesis, the CIFs for place

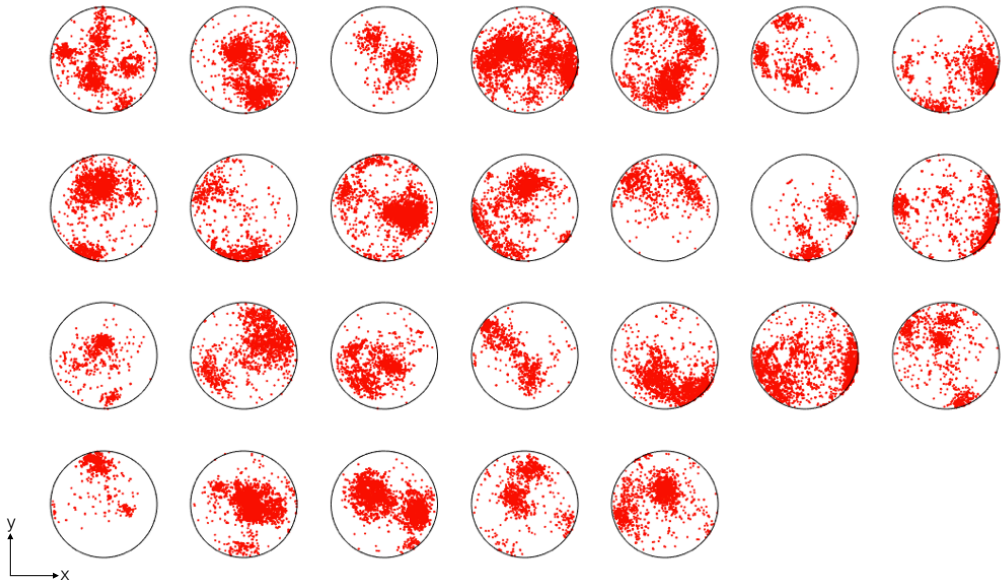

Figure 6.2: Spike scatter plot for multi-modal place and grid cells - Each plot corresponds to spiking activity of a single multi-modal place or grid cell displayed in x, y co-ordinates in the circular arena. Each dot corresponds to x, y coordinates of rat's position when the cell spiked.

and grid cells are modeled as a function of the rat's position (stimuli) and the neuron's spiking history (intrinsic factors), i.e.,

$$\lambda(t|x,y,h) \triangleq \lim_{\Delta t \to 0} \frac{\Pr(\text{spike in time } \Delta t | X{=}x, Y{=}y, \mathcal{H}_t{=}h)}{\Delta t}$$

where $x(t), y(t)$ is the rat's position over time inside an arena, and the vector $\mathcal{H}_t$, consists of spiking history covariates at time t as in (15, 88, 89, 90, 91).

Then the generalized linear model (GLM) and Bayesian estimation approaches are used to estimate the CIF parametrically and non-parametrically, respectively.

6.1.2.1 Generalized linear models

As defined in chapter 1, in GLMs, the logarithm of the CIF is linear in its parameters. Inspired by the success of quadratic and Zernike positional covariates in the modeling of the place cells, these two representations were chosen to estimate the CIFs. In addition, spike history dependence was also modeled with 25 history covariates $h_i; i = 1, \cdots, 25$ that correspond to the number of spikes in $(t - 2i,\ t - 2i + 2)ms$ as done in (15). In particular, the CIF is modeled in following two ways:

1. Gaussian GLM (GLMgauss)

$$\log(\lambda(x, y, \mathcal{H})) = \alpha_1 + \alpha_2 x + \alpha_3 y + \alpha_4 x^2 + \alpha_5 y^2 + \alpha_6 xy + \sum_{i=1}^{25} \beta_i h_i \tag{6.1}$$

2. Zernike GLM (GLMzern)

$$\log(\lambda(x, y, \mathcal{H})) = \sum \gamma_{i,j} \chi(i, j) + \sum_{i=1}^{25} \beta_i h_i, \tag{6.2}$$

where $\alpha_1, \cdots, \alpha_6, \beta_i, \gamma_{i,j}$ are the parameters to be estimated from the data. $\mathcal{H} \triangleq [h_1, \cdots, h_{25}]^T$ and $\chi(i, j)$ are Zernike polynomials of 3rd order. The log-linear assumption on the parameters results in a convex likelihood function which allow a quick estimation of parameters by the parametric ML approach (10, 11).

6.1.2.2 Structure for dependence on history of spiking under Bayesian estimation

Baye's rule (92) allows one to use non-parametric approaches to estimate $\lambda(\cdot)$ as shown below (21):

$$\lambda(t|x,y,h) \simeq \frac{N}{T}\frac{f(x,y,h|\text{spike in time } \Delta t)}{f(x,y,h)} \tag{6.3}$$

where $h(t) \triangleq \log(\text{time since last spike})$, N is the total number of spikes within time interval T, which is the total duration of the spike train observation. Note that the time since last spike is indicated by the symbol τ from here onwards. The use of the logarithm in the history function allows for a smoother dependence of λ on τ, which in turn allows for capturing high frequency components in the CIF due to refractoriness (i.e., sharp decrease in $\lambda(t)$ after a spike) and bursting.

To estimate $f(x,y,h)$ and $f(x,y,h|\text{spike in time } \Delta t)$, Baye's rule is used to decompose the CIF as:

$$\frac{f(x,y,h|\text{spike in time } \Delta t)}{f(x,y,h)} = \frac{f(x,y|\text{spike in time } \Delta t)}{f(x,y)}\frac{f(h|x,y,\text{spike in time } \Delta t)}{f(h|x,y)}. \tag{6.4}$$

Now defining $\lambda_{xy}(x,y) \triangleq \frac{f(x,y|\text{spike in time } \Delta t)}{f(x,y)}$ gives:

$$\frac{N}{T}\frac{f(x,y,h|\text{spike in time } \Delta t)}{f(x,y,h)} = \lambda_{xy}(x,y)\frac{f(h|x,y,\text{spike in time } \Delta t)}{f(h|x,y)} \tag{6.5}$$

$$= \lambda_{xy}(x,y)\frac{N}{T}\frac{f(h|\lambda_{xy}(x,y),\text{spike in time } \Delta t)}{f(h|\lambda_{xy}(x,y))} \tag{6.6}$$

$$= \lambda_{xy}(x,y)\frac{N}{T}\frac{f(h\lambda_{xy}(x,y)|\text{spike in time } \Delta t)}{f(h\lambda_{xy}(x,y))}\times \frac{f(\lambda_{xy})}{f(\lambda_{xy}|\text{spike in time } \Delta t)} \tag{6.7}$$

$$\triangleq \lambda_{xy}(x,y)\lambda_h(h,\lambda_{xy}). \tag{6.8}$$

Above, it is assumed that knowing $\lambda_{xy}(x,y)$ conveys the same information about h as knowing x, y. This assumption is reasonable as h is defined as the log(time since last spike) which depends on the positional co-ordinates x, y through the spiking patterns of neuron which in turn are characterized by $\lambda_{xy}(x,y)$. This assumption reduces a three dimensional density estimation problem into a two dimensional density estimation problem which increases both computational speeds and accuracy (if the assumption is correct). The λ_{xy} and λ_h are both products and ratio of densities which are estimated using both KDE2nd (higher order kernels were too slow for estimation) and `BLMLQuick` methods. Further, to estimate $\lambda_h(h,\lambda_{xy}) \triangleq \frac{N}{T}\frac{f(h\lambda_{xy}(x,y)|\text{spike in time } \Delta t)}{f(h\lambda_{xy}(x,y))}\frac{f(\lambda_{xy})}{f(\lambda_{xy}|\text{spike in time } \Delta t)}$, instead of estimating

$$f(h\lambda_{xy}(x,y)|\text{spike in time } \Delta t), \quad f(h\lambda_{xy}(x,y)), \quad f(\lambda_{xy}|\text{spike in time } \Delta t), \quad f(\lambda_{xy}),$$

$$f(hg(\lambda_{xy}(x,y))|\text{spike in time } \Delta t), \quad f(hg(\lambda_{xy}(x,y))), \quad f(g(\lambda_{xy})|\text{spike in time } \Delta t), \quad f(g(\lambda_{xy})),$$

with $g(\lambda_{xy}) \triangleq \log(0.003 + \lambda_{xy})$ are estimated. This was done because by definition $\lambda_{xy} > 0$ and hence has a non-smooth pdf at $\lambda_{xy} = 0$ which are hard to estimate using non-parametric methods such as KDE and BLML estimation.

The transformation does not effect the estimation λ_h as due to the axioms of probability:

$$\lambda_h(h, \lambda_{xy}(x,y)) \triangleq \frac{N}{T} \frac{f(h\lambda_{xy}(x,y)|\text{spike in time } \Delta t)}{f(h\lambda_{xy}(x,y))} \frac{f(\lambda_{xy})}{f(\lambda_{xy}|\text{spike in time } \Delta t)} \tag{6.9}$$

$$= \frac{N}{T} \frac{f(hg(\lambda_{xy}(x,y))|\text{spike in time } \Delta t)}{f(hg(\lambda_{xy}(x,y)))} \frac{f(g(\lambda_{xy}))}{f(g(\lambda_{xy})|\text{spike in time } \Delta t)} \tag{6.10}$$

for any one-to-one transformation $g(\cdot)$.

6.1.3 Model construction and selection

The goodness-of-fit of the four estimates of λ are computed using the time rescaling theorem and the Kolmogorov-Smirnov (KS) statistic (93). Briefly, 80% of the data is used to estimate λ and then the empirical CDF of rescaled spike times is computed using the remaining 20% test data, which should follow a uniform CDF if the estimate of λ is accurate. The similarity between the two CDFs is quantified using the normalized KS-statistic and visualized using the KS-plot (93). A value of KS$>$ 1 indicates that the estimated λ is too extreme ($p < 0.05$) to generate the test data. The closer the normalized KS-statistic is to 0, the better is the estimate.

Although, place and grid cell fields have different shapes and locations, it is hypothesized that they may have the same smoothness across the entire population, i.e. one vector

of bandwidths or cut-off frequencies may work well for the entire population of neurons. To select this vector of bandwidths and cut-off frequencies, performance as measured by goodness-of-fit on test data is evaluated as done in (21). In particular, these bandwidths and cut-off frequencies are chosen after testing their different combinations. The frequencies and bandwidths that give the lowest decoding error on test data (see section 6.2.2) are selected. The selected values for bandwidths and cut-off frequencies are $q_x = q_y = 0.143n^{-0.2}, q_h = 0.191n^{-0.2}, q_\lambda = 0.222n^{-0.2}$ and $f_{cx} = f_{cy} = 4.4, f_{ch} = 4.0, f_{c\lambda} = 2.4$ for the KDE2nd and `BLMLQuick` methods respectively.

6.2 Results

6.2.1 CIF estimation

Figure 6.3B and 6.4B, show KS-plots (93) for the `BLMLQuick`, KDE2nd, GLMgauss and GLMzern estimators. It is clear from the figures, that both the non-parametric estimates performed better than both non-parametric estimates as their KS-plots are closest to 45 degree line. In particular, the KS-statistics values for `BLMLQuick`, KDE2nd, GLM Gaussian and GLM Zernike methods were $(0.54, 0.64)$, $(0.89, 0.61)$, $(2.49, 3.10)$ and $(4.07, 4.49)$, for the place and grid cells shown in 6.3 and 6.4 respectively. On entire population of uni-modal place cells, multi-modal place cells and grid cells BLML methods (average KS-statistics= 1.45) did significantly ($p = 0.008$,$p = 1.69E-14$, $p = 1.38E-20$ paired t-test respectively) better than the KDE methods (average KS-statistics= 1.72), GLM gaussian (average KS-statistics= 3.18) and GLM zernike methods (average KS-statistics = 4.44).

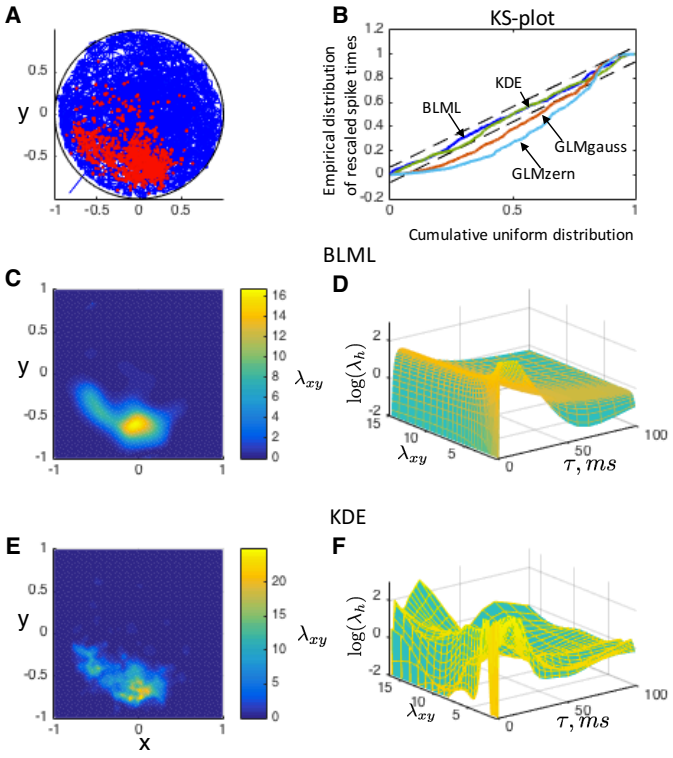

Figure 6.3: Comparision of `BLMLQuick`, KDE and GLM methods for a complex Place cell - (**A**) The trajectory of rat during the duration of the experiment is shown by the blue line. Each dot marks (x, y) co-ordinates of rat's position when the place cell spikes. (**B**) The KS-plot for `BLMLQuick` (blue), KDE2nd (green), GLM Gaussian (red) and GLM Zernike (light blue), along with 95% confidence intervals (dashed ines) using 20% of test data. (**C,E**) Estimated $\lambda_{xy}(x, y)$ for `BLMLQuick` and KDE2nd under Bayesian framework respectively. (**D,F**) Estimated $\lambda_h(h, \lambda_{xy})$ for `BLMLQuick` and KDE2nd under Bayesian framework respectively.

Figure 6.3C, E, and 6.4C, E, plot estimates for $\lambda_{xy}(x, y)$ using `BLMLQuick` and KDE2nd, respectively. First of all it can be seen that for both place and grid cells, $\lambda_{xy}(x, y)$ is larger wherever the neuron spikes more, verifying the "place-like" and "grid-like" behaviour. Second, the `BLMLQuick` method produces a smoother $\lambda_{xy}(x, y)$ than KDE2nd methods.

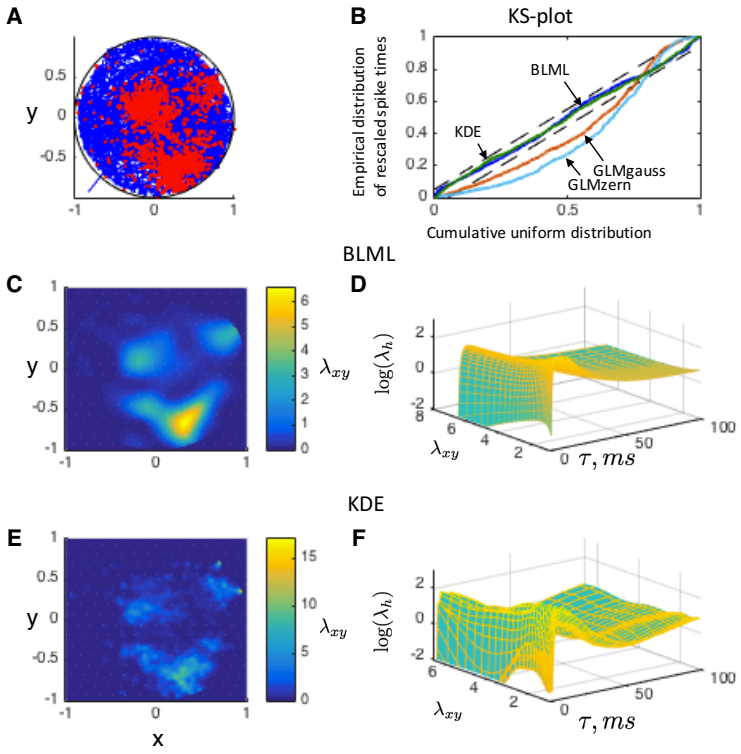

Figure 6.4: Comparision of `BL-MLQuick`, KDE and GLM methods for a complex Grid cell - (**A**) The trajectory of rat during the duration of the experiment is shown by blue line. Each dot marks (x,y) co-ordinates of rat's position when the place cell spikes. (**B**) The KS-plot for `BLMLQuick` (blue), KDE2nd (green), GLM Gaussian (red) and GLM Zernike (light blue), along with 95% confidence intervals (dashed ines) using 20% of test data. (**C,E**) Estimated $\lambda_{xy}(x, y)$ for `BLMLQuick` and KDE2nd under Bayesian framework respectively. (**D,F**) Estimated $\lambda_h(h, \lambda_{xy})$ for `BLMLQuick` and KDE2nd under Bayesian framework respectively.

Figure 6.3D, F and 6.4D, F plot $\lambda_h(h, \lambda_{xy})$ for the `BLMLQuick` and KDE2nd methods for a place and grid cell, respectively. Both methods are successful in capturing the known behaviour of refractoriness and bursting in the neuronal activity.

Refractoriness is captured by smaller values near $\tau = 0$ across all λ_{xy}, and burst-

ing is captured by sudden large values around $\tau = 4ms$, across all λ_{xy}. The values achieved at $\tau = 4$ and smaller λ_{xy} are larger than the values achieved at $\tau = 4ms$ and larger λ_{xy}. Looking holistically this suggests that even if λ_{xy} is small, $\lambda(t)$ ($\lambda(t) = \lambda_{xy}(x(t), y(t))\lambda_h(h(t), \lambda_{xy}(x(t), y(t)))$) can still reach high values if there is a spike $4ms$ in the past. This, in turn, makes the values of $\lambda(t)$ after $4ms$ of any spike almost constant and hence independent of x, y. These findings suggest that bursting in place and grid cells is an *internal* phenomenon that is not affected much by the external stimulus. More importantly assuming the history dependence allowed to capture the extra variance (2, 87) previously been observed in the spiking activity of place cells that is not accounted by inhomogeneous poisson process models that only depends on x and y cordinates of rats position. In particular, the place cell shown in figure 6.3 shows a KS-statistics of 0.54 with covariates for history dependece and a KS-statistics of 5.1 without the history covariates.

Figure 6.5 and 6.6 plot the recorded and simulated activity (using BLML model) for the place and grid cells. It can be seen that both at the macro and micro level, the BLML model was able to reproduce the recorded activity, i.e., the x and y dependence is nicely captured as shown in figure 6.5B and 6.6B along with the micro level bursting as shown in seen in figure 6.5D and 6.6D for both the place and grid cells. Further, it can be seen in the zoomed in view in figure 6.5C and 6.6C that there is high variability (2, 87) between the spike counts on different trajectories that pass through the zoomed in region, even though the λ_{xy} is almost constant over this small region. As shown by the goodness of fit previously this variability can safely be accounted to the dependence on history, particularly bursting,

which makes the neuron spikes atleast 3-4 times if it spikes once, the neuron does not spike at all otherwise. This phenomenon is again observed in the simulated activity shown in 6.5D and 6.6D again illustrating the success of model in capturing this variability.

6.2.2 Decoding

After selecting the Bayesian methods over GLM models to represent the complex place fields and grid fields, decoding is performed. To decode the rat's position from neuronal activity, maximum a posteriori (MAP) estimation is implemented. First, the likelihood of observing h_i, $i = 1, \cdots, m$ (m being the number of neurons) given the rat's positional co-ordinates x, y is calculated as follows:

$$L(h_1, \cdots, h_m|x, y) = \prod_{i=1}^{m} f(h_i|x, y) = \prod_{i=1}^{m} \frac{f(x, y, h_i)}{f(x, y)}. \tag{6.11}$$

It is assumed that the neurons fire conditionally independently such that $f(h_i, h_j|xy) = f(h_i|x, y)f(h_j|x, y)$ for all i, j. This allowed the calculation of the MAP estimator as follows:

$$(\hat{x}, \hat{y}) = \text{argmax}_{x,y}(L(h_1, \cdots, h_m|x, y)f(xy)). \tag{6.12}$$

The density $f(x, y, h_i)$ is a three dimensional pdf which may be difficult to estimate directly. Hence, it is estimated assuming that λ_{xy} is a sufficient statistic (36) for h (that is, knowing x, y does not convey any additional information about h if $\lambda_{xy}(x, y)$ is known) which converts it into product and ratio of two dimensional pdfs as shown below.

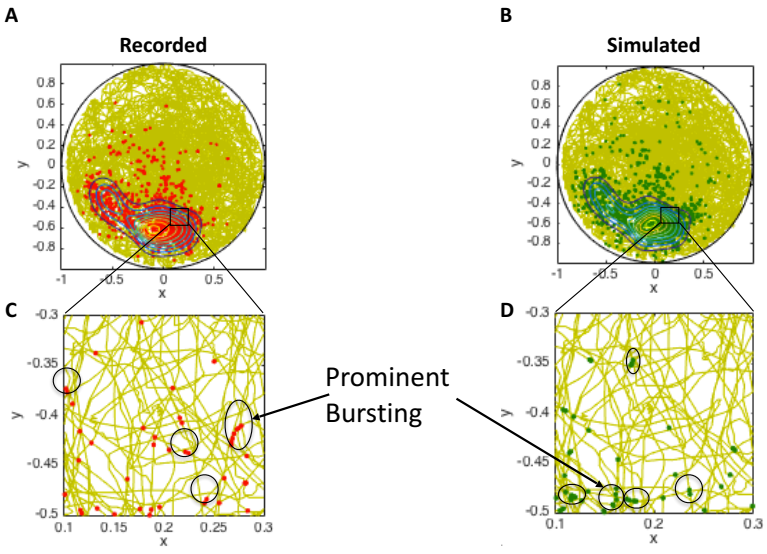

Figure 6.5: Simulated activity for a complex place cell using `BLMLQuick` model - (**A,C**) The recorded spiking activity of the complex place cell shown in figure 6.3 and its zoom in view. (**B,D**)) The simulated spiking activity of a place cell and its zoom in view. Each dot marks (x,y) co-ordinates of rat's position when the place cell spikes.

$$f(x, y, h_i) = f(x, y)f(h_i|\lambda_{xy}(x, y)) = f(x, y)\frac{f(h_i\lambda_{xy})}{f(\lambda_{xy})} = f(x, y)\frac{f(h_i g(\lambda_{xy}))}{f(g(\lambda_{xy}))}, \qquad (6.13)$$

where $g(\lambda_{xy}) = log(0.003+\lambda_{xy})$ is a one-to-one function. The densities $f(x, y)$, $f(h_i g(\lambda_{xy}))$ and $f(g(\lambda_{xy}))$ are estimated by either the KDE2nd or `BLMLQuick` methods with bandwidth or cut-off frequencies equal to the ones found in previous section.

Figure 6.7A, B, C plot the reconstructed $\hat{x}(t)$ and $\hat{y}(t)$ (calculated using `BLMLQuick`) for a six minute period using test data from only unimodal place cells, only multimodal place and grid cells (shown in figure 6.1 and 6.2) and both respectively. Overlaid on top of $\hat{x}(t)$

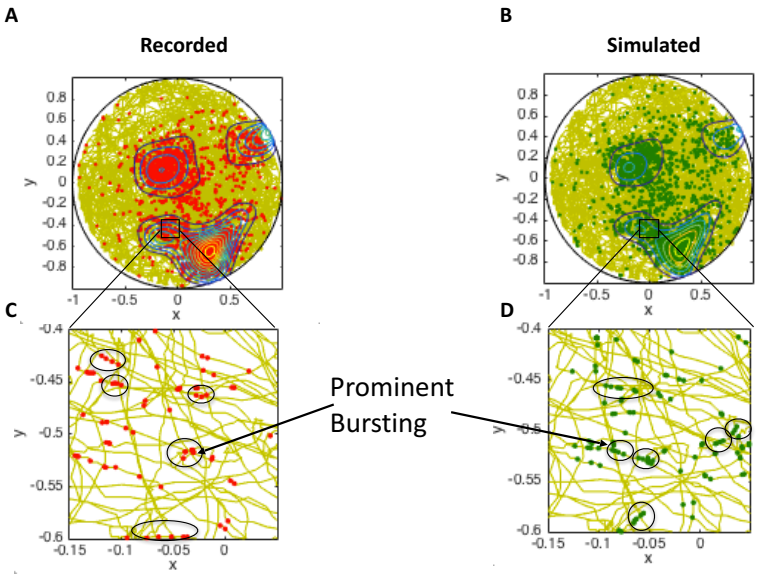

Figure 6.6: Simulated activity for a grid cell using `BLMLQuick` model - (**A,C**) The recorded spiking activity of the grid cell shown in figure 6.4 and its zoom in view. (**B,D**)) The simulated spiking activity of the grid cell and its zoom in view. Each dot marks (x,y) co-ordinates of rat's position when the place cell spikes.

and $\hat{y}(t)$ values are the recorded $x(t)$ and $y(t)$ of rat during the same time period. The $r^2 = 0.83, 0.85$ and 0.89 for the reconstruction with place cell only, grid cell only and both combined, which is marginally superior to the $r^2 = 0.81, 0.82, 0.89$ values obtained by using KDE2nd method (instead of `BLMLQuick`) for the reconstruction with place cell only, grid cell only and both combined, respectively. The comparable performance of `BLMLQuick` and KDE may be due to the way bias-variance trade-off is handled by `BLMLQuick` and KDE2nd methods. Particularly, adaptive KDE automatically reduces bias of the estimate when the number of data samples is high, similarly adaptive methods have not yet been developed

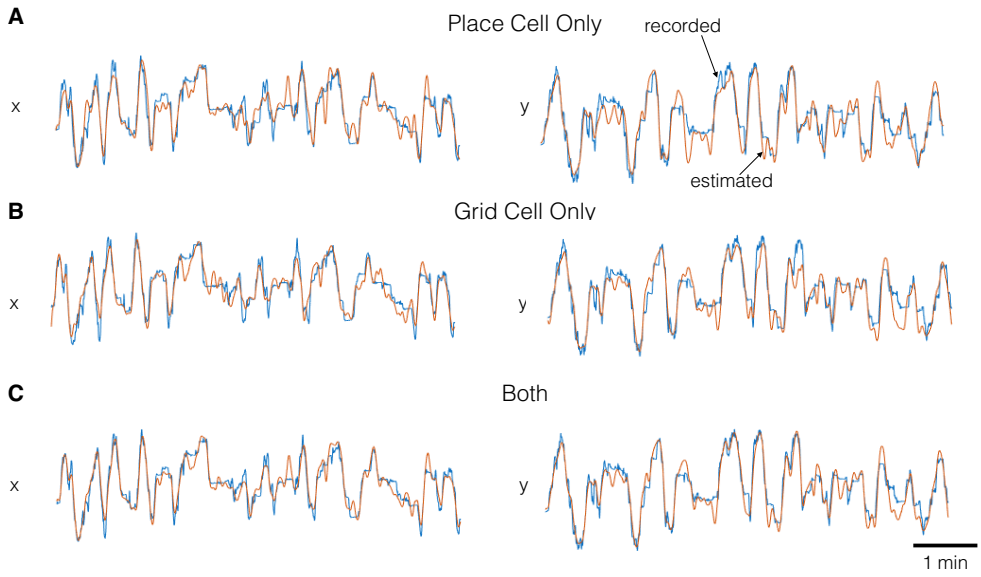

Figure 6.7: Reconstructed rats trajectory using `BLMLQuick` - (**A,B,C**) The estimated $\hat{x}(t)$, $\hat{y}(t)$ for a 6 minute test period using the place cell only, grid cell only and both cells respectively. Overlaid on top are the recorded $x(t)$, $y(t)$ trajectory of the rat during the same period.

for BLML methods. On the other hand, BLML being a maximum likelihood estimator produces estimators with less variance as compared to KDE.

6.3 Conclusion

In this chapter, the "fourier hypothesis" (22) is first tested. It states that the fields of place cells form by fourier like summation of the fields of grid cells. As BLML estimator also assumes that the underlying densities are sum of sinusoidal functions, it is a natural cadidate to test the fourier hypothesis. To do this the BLML estimator is used for building encoding and decoding models for spiking activity for uni-modal and multi-modal place cells,

and grid cells. The performance of these models is then compared with that of parametric GLM models (using both quadratic or Zernike polynomials) and non-parametric KDE (2nd order Gaussian kernels) methods. The BLML estimator outperforms both the parametric methods. The performance of KDE methods is found comparable with BLML however the difference becomes significant when accumulated over the entire population of recorded neurons. The better performance of the BLML estimator supports the "fourier hypothesis".

The encoding models presented in this chapter are able to capture the neuronal spiking at a millisecond resolution. This allowed to capture the dependence of spiking on history of spiking. Particularly, the BLML models capture the contribution of bursting and refractoriness to spiking of multi-modal place and grid cells, which to the authors knowledge has not been achieved previously. This in turn explain the extra variance which have been reported previously in the firing patterns of place cells and can not able to be accounted by inhomogeneous poisson process models that are dependent only on x, y co-ordinates of the rat's position (2, 87).

Further, due to the non-parametric nature of BLML estimator, the same model structure (with the same cut-off frequency) is able to characterize a variety of field types observed in both place and grid cells. This flexibility is really helpful and is a step towards a practical solution for robust decoding via an implantable device. Currently, decoding via implantable devices suffer from drift in recordings and neuronal firing over time. Such drifts may include change in firing patterns of the cell itself e.g. grid and place cells are known to reorient their fields over time (87, 94, 95) or change in firing because of change in neuronal structure and

orientation and/or electrode position. Therefore, to achieve robust decoding the encoding models should be flexible so that they can model a variety of field types. As shown in this chapter, the non-parametric methods such as BLML are flexible enough to model a variety of field types that can be generated by such drifts accurately and therefore the decoding performance of the devices using BLML (or possibly KDE methods) may remain robust over time.

More importantly, the decoding results shown in this chapter are the first attempt to decode trajectory of rat in a two dimensional space using multi-modal place cell and grid cell fields. As shown in this chapter the multi-model place cells and grid cells are difficult to characterize by state-of-the-art parametric models. Such models have shown success only in decoding from uni-modal place cell fields (14, 93). The mixture of Gaussian model may show some success in characterizing multi-modal fields but these models are computationally very slow making their implementation problematic (96). The non-parametric models have shown some success in decoding from multi-unit activity and/or multi-modal fields but decoding using such models have mainly been limited to one dimensional setting (21).

Chapter 7

Discussion and Future Work

In this chapter various possible extensions and uses of the BLML estimator and BLML theory are discussed. These extensions may have direct applications in the field of machine learning and imaging science.

7.1 Developing BLML theory further

The BLML estimator is introduced in this thesis is the first demonstration of the existence of a non-parametric maximum likelihood estimator with a square-root representation. Although, some properties of the BLML estimator have been proved in this thesis, there remain other properties which need to be established. These properties are outlined below.

7.1.1 Asymptotic analysis

This thesis establishes the consistency of the BLML estimator when the true density $f(x) > 0$, however, a proof for the consistency when $f(x) \geq 0$ is still needed. Further, parametric

ML estimators are known to achieve the Cramer-Rao lower bound asymptotically, however, it is unclear if the non-parametric ML estimators also show an analogous property. To show this, first the Cramer-Rao bound needs to be generalized to the non-parametric setting and then the efficiency of BLML methods must be computed. Finally, parametric ML methods show a convergence rate of $\mathcal{O}\left(\frac{1}{n}\right)$, and although under simulations BLML methods show similar convergence rates, their convergence rates need to be established in theory.

7.1.2 Generalization to other orthonormal systems

BLML estimators are a particular kind of orthogonal transform square-root density estimators that use the Fourier basis. However, the proofs are not dependent on the choice of the basis. More generally, many orthonormal basis systems must result in a corresponding ML estimator. The author conjectures that, in general, estimators of the form

$$\hat{f}(x) = \left(\frac{1}{n} \sum_{i=1}^{n} \hat{c}_i K_h(x - x_i) \right)^2 \tag{7.1}$$

where $K(\cdot)$ is a radial kernel, h is the bandwidth of kernel, $\hat{\mathbf{c}} \triangleq [\hat{c}_1, \cdots, \hat{c}_n]^T$ and

$$\hat{\mathbf{c}} = \underset{\boldsymbol{\rho_n}(\mathbf{c})=\mathbf{0}}{\arg\max} \left(\prod_{i=1}^{n} \frac{1}{c_i^2} \right). \tag{7.2}$$

Here $\rho_{ni}(\mathbf{c}) \triangleq \frac{1}{n} \sum_{j=1}^{n} c_j k_{ij} - \frac{1}{c_i} \ \forall \ i = 1, \cdots, n$ and

$s_{ij} \triangleq K_h(x_i - x_j) \quad \forall \ i, j = 1, \cdots, n$

may also be non-parametric ML estimators.

7.1.3 Algorithms for exact BLML estimator

This thesis proposes three algorithms to compute the BLML estimator. Out of these three algorithms, `BLMLTrivial` gives the exact BLML estimator asymptotically if $f(x) > 0$. In practice, these two assumptions hold most of the time (sample sizes are as small as $n > 100$ give asymptotic effects and pdfs in nature are mostly strictly greater than zero). Nonetheless, there remains a need to further improve the algorithms so that they yield the exact BLML estimator even in the cases where $f(x) \geq 0$ and when only a small number of samples are available.

7.2 Binary classification using the BLML estimator

The BLML estimator may also be very useful for binary classification with hard boundaries. Classifiers look for hard boundaries (possibly smooth) that divide the space of random variables into two classes. A hard boundary (as opposes to a soft boundary) means that each outcome belongs to one of the two classes with certainty (as oppose to probabilistically). If such a boundary exist, then the probability of observing an outcome at the boundary should be zero because otherwise there will eventually be an outcome at the boundary violating the assumption of its existence. Now, if $f(\mathbf{x}) = g^2(\mathbf{x})$ is a band-limited pdf that denotes the probability of observing an outcome at point $\mathbf{x}$ in the space of outcomes and $g(x)$ is the corresponding band-limited square-root of $f(x)$, then both $f(x)$ and $g(x)$ are zero at the boundary. This would further imply that $g(x) > 0$ in the space of one class of outcomes and $g(x) < 0$ in the space of the other class (due to band-limited assumption

on $g(x)$, for detail reasoning see Theorem B.7.4). Now, consider the problem of supervised classification as discussed below.

7.2.1 Supervised classification

In supervised classification (97), the labels of the class where the observed data points belong are known for a training data set. Therefore, without loss of generality, it can be assumed that $g(\mathbf{x}_i^{(a)}) > 0$ for all training samples in class (a) and $g(\mathbf{x}_i^{(e)} < 0)$ for all training samples in class (e). Now, the problem of classification reduces to finding a $g(\mathbf{x})$ such that $g^2(x)$ is a pdf, $g(\mathbf{x}_i^{(a)}) > 0$, and $g(\mathbf{x}_i^{(e)}) < 0$. Many g's exists that satisfy these properties. Out of all such g's, the functions that can describe the training data optimally and result in smoother boundaries $g(x) = 0$ are desirable. These two properties are directly addressed by first assuming that $g(x)$ and hence $g^2(x)$ are band-limited, i.e.,

$$g^2(\mathbf{x}) \in \mathbb{U}(\omega_c). \tag{7.3}$$

where, $\mathbb{U}(\omega_c)$ is set of band limited pdfs (see appendix A for details) and $g^2(x)$ is a maximum likelihood estimate on training data. Adding these two assumptions, the problem of classification can be written as the following optimization problem:

$$\tilde{g}(\mathbf{x}) = \underset{g^2 \in \mathbb{U}(\omega_c), g(\mathbf{x}_i^{(a)})>0, g(\mathbf{x}_i^{(e)})<0}{\arg\max} \prod_{i=1}^{n_a} g^2\left(\mathbf{x}_i^{(a)}\right) \prod_{i=1}^{n_e} g^2\left(\mathbf{x}_i^{(e)}\right) \tag{7.4}$$

here n_a and n_e are the number of elements in classes (a) and (e) in the training data set, respectively. Using BLML theory, the solution to the above problem is as follows:

$$\tilde{g}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n_a} \tilde{c}_i^{(a)} \operatorname{sinc}_{\mathbf{f}_c}(\mathbf{x} - \mathbf{x}_i^{(a)}) + \frac{1}{n}\sum_{i=1}^{n_e} \tilde{c}_i^{(e)} \operatorname{sinc}_{\mathbf{f}_c}(\mathbf{x} - \mathbf{x}_i^{(e)}) \tag{7.5}$$

where $n = n_a + n_e$ is the total number of samples, $\tilde{\mathbf{c}} = [c_1^{(a)}, \cdots, c_{n_a}^{(a)}, c_1^{(e)}, \cdots, c_{n_e}^{(e)}]^T$ is the solution of $\rho_n(c) = 0$, in the orthant with orthant indicator vector $c_0 = [\mathbf{1}(n_a), -\mathbf{1}(n_e)]$ where $\mathbf{1}(n)$ represents a vectors of ones with n elements. The matrix $\mathbf{S}$ (see) formed by elements s_{ij} in the definition of ρ is arranged according to the elements of two classes, with the upper left part containing elements only from class (a), the lower right part containing elements from only class (e), and the upper right and lower left parts containing cross terms.

Once $\tilde{g}(\mathbf{x})$ is estimated, its sign can be evaluated to classify a test sample in one step. The classification boundary is given by equation $\tilde{g}(\mathbf{x}) = 0$. Note that the accent $(\tilde{\cdot})$ is used here to distinguish it from $(\hat{\cdot})$ which indicates a global maximum (over all orthants) BLML solution as opposed to the maximum in any given orthant.

Figure 7.1 plots the classification results using the above BLML classification algorithm on surrogate data with the true boundary being a circle at center 0 and radius 1. It can be seen that the BLML classification algorithm successfully identifies the true boundary in this example. A more complete study for the properties of the BLML classification and its comparison with standard classification approaches needs to be done.

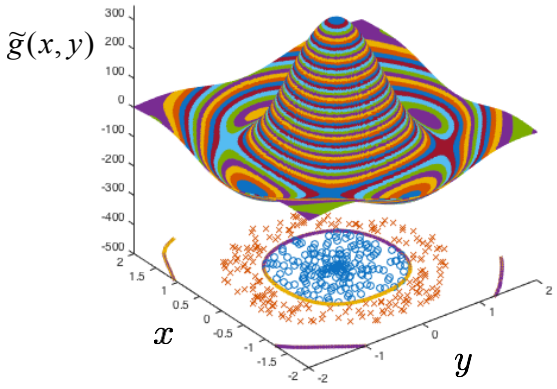

Figure 7.1: Classification using BLML method - Illustrating the training samples and their classes a (circles) and b (crosses). Estimated $\tilde{g}(x, y)$ is shown on top of these samples and the estimated classification boundary is drawn by gradient colored line (blue-yellow).

7.2.1.1 Mislabeled data

Mislabelled data can also be handled using BLML methods. One strategy is to first choose the orthant indicated by the labels as the base orthant. Then computing $\tilde{g}(\mathbf{x})$ in the base orthant and all $n_a + n_b$ orthants neighbouring (hamming distance one between orthant indicator vectors) the base orthant and subsequently computing the corresponding $1 + n_a + n_b$ log likelihood values $\sum \log(\tilde{g}^2(\mathbf{x}_i^{(a)})) + \sum \log(\tilde{g}^2(\mathbf{x}_i^{(b)}))$. These values are log-likelihood of observing the training data assuming that at max one of the labels in base orthant is incorrect. Therefore, the next step is to choose the orthant that gives the maximum likelihood value among all $1 + n_a + n_b$ orthants. If the orthant that maximizes the likelihood is the base orthant, stop otherwise set the orthant that maximizes the likelihood as base

orthant and repeat the process. After running this algorithm the labels corresponding to the final orthant would be the corrected labels.

7.2.2 Unsupervised binary classification

The previous section shows that the problem of binary classification is inherently related with finding the correct orthant for the BLML estimator. In supervised learning, this orthant is given by the labels of the data points. In unsupervised learning (98) these labels are not provided. The BLML can be useful in such cases too as the correct orthant can directly be determined as the orthant that maximizes likelihood globally (over all orthants). However, as discussed in chapter 4 this problem is exponentially hard and heuristic and slow algorithms such as `BLMLBQP` have to be used to obtain the correct orthant. Making these algorithms exact and improving their computational speed is left for future study.

7.3 Application to image processing

Images, in general, can be viewed as a scaled two dimensional pdf. The intensity values at each pixel may correspond to the histogram generated by collecting samples from such a pdf. The field of image processing focuses on reducing the noise in such histograms and thereby finding important features from imaging data. BLML inherently is a procedure of smoothing the histogram while maximizing the likelihood. Therefore, there may be direct applications of the BLML estimator in image processing, particularly less exposure imaging, including high shutter speed, low light and microscopic imaging. The BLML estimator, due

to its relation to wave functions (see chapter 4 discussion), may also be useful in x-ray crystallography, imaging the microlevel slit structure using diffraction patterns, and more.

Appendix A

BLML Estimator

A.1 Preliminaries and formulation of the BLML estimator

Consider a pdf, $f(x)$, of a random variable $x \in \mathbb{R}$ with Fourier transform $F(\omega) \triangleq \int f(x)e^{-i\omega x} dx$. Let $\mathbb{U}(\omega_c)$ be the set of band-limited pdfs with frequency support in $(-\omega_c, \omega_c)$, i.e., $\omega_c \in \mathbb{R}$ is the cut-off frequency of the Fourier transform of the pdf. Then,

$$\mathbb{U}(\omega_c) = \left\{ f : \mathbb{R} \to \mathbb{R}^+ \mid \int f(x)dx = 1, F(\omega) = 0 \ \forall |\omega| > \omega_c \right\} \tag{A.1}$$

Since each element $f \in \mathbb{U}(\omega_c)$ is a pdf, it can be written as $f(x) = g^2(x)$, where $g \in \mathbb{V}(\omega_c)$ is defined as:

$$\mathbb{V}(\omega_c) = \left\{ g : \mathbb{R} \to \mathbb{R} \mid g(x) = \sqrt{f(x)},\ f \in \mathbb{U}(\omega_c) \right\} \tag{A.2}$$

Finally, $\mathbb{W}(\omega_c)$ can be defined as the set of all Fourier transforms of elements in $\mathbb{V}(\omega_c)$:

$$\mathbb{W}(\omega_c) = \left\{ G : \mathbb{R} \to \mathbb{C} \mid G(\omega) = \int g(x)e^{-i\omega x} dx,\ g \in \mathbb{V}(\omega_c) \right\} \tag{A.3}$$

Note that since $f(x) \in \mathbb{U}(\omega_c)$ is band limited, $g(x) \in \mathbb{V}(\omega_c)$ will also be band limited in $(\frac{-\omega_c}{2}, \frac{\omega_c}{2})$. Therefore, $G(\omega) = 0 \ \forall |\omega| > \frac{\omega_c}{2} \ \forall G \in \mathbb{W}(\omega_c)$. Finally, $\mathbb{V}(\omega_c)$ and $\mathbb{W}(\omega_c)$ are Hilbert spaces with the inner product defined as $< a, b >= \int a(x)b^*(x)dx$, $< a, b >= \frac{1}{2\pi} \int a(\omega)b^*(\omega)d\omega$, respectively. The norm $||a||_2^2 =< a, a >$ is defined for both spaces. Further, note that for all elements in $\mathbb{V}(\omega_c)$ and $\mathbb{W}(\omega_c)$, $||a||_2^2 =< a, a >= 1$.

The likelihood function for band-limited pdfs Now consider a random variable, $x \in \mathbb{R}$, with unknown pdf $f(x) \in \mathbb{U}(\omega_c)$ and its n independent realizations $x_1, x_2, \cdots, x_n$. The likelihood $L(x_1, \cdots, x_n)$ of observing $x_1, \cdots, x_n$ is then:

$$L(x_1, \cdots, x_n) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} g^2(x_i), \ g \in \mathbb{V}(\omega_c) \tag{A.4a}$$

$$= \prod_{i=1}^{n} \left(\frac{1}{2\pi} \int G(\omega) e^{j\omega x_i} d\omega \right)^2, \ G \in \mathbb{W}(\omega_c) \tag{A.4b}$$

Defining:

$$b_i(\omega) \triangleq \left\{ \begin{array}{cc} e^{-j\omega x_i} & \forall \ \omega \in (-\frac{\omega_c}{2}, \frac{\omega_c}{2}) \\ 0 & o.w. \end{array} \right\} \tag{A.5}$$

gives:

$$L(x_1, \cdots, x_n) = \prod_{i=1}^{n} (< G(\omega), b_i(\omega) >)^2 \triangleq L[G]. \tag{A.6}$$

Further, consider $\hat{G}(\omega)$ which maximizes the likelihood function:

$$\hat{G} = \underset{G \in \mathbb{W}(\omega_c)}{\arg\max} (L[G]). \tag{A.7}$$

Then the BLML estimator is:

$$\hat{f}(x) = \left(\frac{1}{2\pi} \int \hat{G}(\omega) e^{j\omega x} d\omega \right)^2. \tag{A.8}$$

A.2 Proof of theorem 4.1.1

Proof: Because of (A.5), (A.7) is equivalent to

$$\hat{G}(\omega) = \underset{G:\mathbb{R}\to\mathbb{C}, ||G||_2^2=1}{\arg\max} (L[G]). \tag{A.9}$$

Note that Parseval's equality (99) is applied to get the constraint $||G||_2^2 = 1$. Now, the Lagrange multiplier (100) is used to convert (A.9) into the following unconstrained problem:

$$\hat{G}(\omega) = \underset{G:\mathbb{R}\to\mathbb{C}}{\arg\max} \left(L[G] + \lambda \left(1 - ||G||_2^2\right)\right). \tag{A.10}$$

$\hat{G}(\omega)$ can be computed by differentiating the above equation with respect to G using calculus of variations (101) and equating it to zero. This gives:

$$\hat{G}(\omega) = \frac{1}{n} \sum_{i=1}^{n} c_i b_i(\omega) \tag{A.11a}$$

$$c_i = \frac{n}{\lambda} \prod_{j \neq i} \left(< \hat{G}(\omega), b_j(\omega) > \right)^2 < \hat{G}(\omega), b_i(\omega) > \quad \text{for} \quad i = 1 \cdots n \tag{A.11b}$$

To solve for c_i, the value of $\hat{G}$ is substituted back from (A.11)a into (A.11)b and both

sides are multiplied by $< \hat{G}(\omega), b_i(\omega) >$ to get:

$$c_i \sum_{j=1}^{n} c_j < b_j(\omega), b_i(\omega) >= n^2 k \text{ for } i = 1 \cdots n, \tag{A.12a}$$

$$k \triangleq \frac{1}{n^{2n}\lambda} \left(\prod_{j=1}^{n} \left(\sum_{i=1}^{n} c_i < b_i(\omega), b_j(\omega) > \right)^2 \right) \tag{A.12b}$$

$$= \frac{1}{n^{2n}\lambda} \left(\prod_{j=1}^{n} \left(\sum_{i=1}^{n} c_i s_{ij} \right)^2 \right) \tag{A.12c}$$

To go from (A.12)b to (A.12)c, observe that $< b_i(\omega), b_j(\omega) > = \frac{\sin(\pi f_c(x_i - x_j))}{\pi(x_i - x_j)} = s_{ij}$ (here $f_c = \frac{\omega_c}{2\pi}$). Now by defining,

$$\mathbf{S} \triangleq \begin{bmatrix} s_{11} & \cdots & s_{1n} \\ \vdots & \ddots & \vdots \\ s_{n1} & \cdots & s_{nn} \end{bmatrix}, \tag{A.13}$$

and using (A.11)a and the constraint $||\hat{G}(\omega)||_2^2 = 1$, one can show that $\mathbf{c}^T\mathbf{S}\mathbf{c} = n^2$. Also, summing up all n constraints in (A.12)a gives $\mathbf{c}^T\mathbf{S}\mathbf{c} = n^3 k$, hence $k = 1/n$. Now, substituting the value of k into (A.12)a and rearranging terms gives the following n constraints:

$$\frac{1}{n} \sum_{j=1}^{n} c_j s_{ij} - \frac{1}{c_i} = \rho_{ni}(\mathbf{c}) = 0 \quad \text{for} \quad i = 1 \cdots n. \tag{A.14}$$

As mentioned in the chapter 4, the above system of equations ($\boldsymbol{\rho_n}(\mathbf{c}) = \mathbf{0}$) is monotonic, i.e., $\frac{\mathrm{d}\boldsymbol{\rho}_\mathbf{n}}{\mathrm{d}\mathbf{c}} > \mathbf{0}$, but with discontinuities at each $c_i = 0$. Therefore, there are 2^n solutions, with each solution located in each orthant, identified by the orthant vector $\mathbf{c}_0 \triangleq sign(\mathbf{c})$. Each of these solutions can be found efficiently by choosing a starting point in a given orthant and

applying numerical methods from convex optimization theory to solve for (A.14). Thus, each of these 2^n solutions corresponds to a local maximum of the likelihood functional $L[G]$. The global maximum of $L[G]$ can then be found by evaluating the likelihood for each solution $\mathbf{c} = [c_1, \cdots, c_n]^T$ of (A.14). The likelihood value at each local maximum can be computed efficiently by using the following expression:

$$L(\mathbf{c}) = \prod_i \left(\frac{1}{n} \sum_j c_j s_{ij} \right)^2 = \prod_i \frac{1}{c_i^2}. \tag{A.15}$$

This expression is derived by substituting (A.11)a into (A.6) and then substituting (A.14) into the result. Now the global maximum $\hat{\mathbf{c}}$ can be found by solving (7.2). Once the global maximum $\hat{\mathbf{c}}$ is computed, we can put together (A.5),(A.8) and (A.11)a to write our solution as (4.1).

Appendix B

Consistency of the BLML estimator

B.1 Bounds on bandlimited PDF

In this section the following theorem is first stated and proved.

Theorem B.1.1 *For all* $f \in \mathbb{U}(\omega_c)$ $f(x) \leq \frac{\omega_c}{2\pi}$ $\forall x \in \mathbb{R}$.

Proof: Above theorem can be proven by finding:

$$y = \max_{f \in \mathbb{U}(\omega_c)} \max_{x \in \mathbb{R}} f(x). \tag{B.1}$$

Because a shift in the pdf domain (e.g. $f(x - \mu)$) does not change the magnitude or bandwidth of $F(\omega)$, without loss of generality one can assume that $\max_{x \in \mathbb{R}} f(x) = f(0)$

and write the above equation as

$$y = \max_{f \in \mathbb{U}(\omega_c)} (f(0)) \tag{B.2a}$$

$$= \max_{g \in \mathbb{V}(\omega_c)} ((g^2(0)) \tag{B.2b}$$

$$= \max_{G \in \mathbb{W}(\omega_c)} \left(\left(\int_{-\omega_c}^{\omega_c} G(\omega) \mathrm{d}\omega \right)^2 \right) \tag{B.2c}$$

$$= \max_{||G||^2 = 1} \left(\left(\int_{-\infty}^{\infty} G(\omega) b(\omega) \mathrm{d}\omega \right)^2 \right) \tag{B.2d}$$

$$= \max \left(\left(\int G(\omega) b(\omega) \mathrm{d}\omega \right)^2 + \lambda(||G||^2 - 1) \right) \tag{B.2e}$$

Here $b(\omega) = 1 \iff |\omega| < \pi f_c$ and is 0 otherwise. Now by differentiating (B.2)e and subsequently setting the result equal to 0, gives $G^*(\omega) = \frac{b(\omega)}{\sqrt{f_c}}$. Therefore $g^*(x) = \frac{\sin(\pi f_c x)}{\pi \sqrt{f_c} x}$, which gives $y = f_c = \frac{\omega_c}{2\pi}$.

Corollary: By the definition of $\mathbb{V}(\omega_c)$, one can apply Theorem B.1.1 and show that for all $g \in \mathbb{V}(\omega_c),\ g(x) \leq \sqrt{\frac{\omega_c}{2\pi}}$.

B.2 Sequence $\bar{c}_{nj}$

Now a sequence $\bar{c}_{nj}$ is defined and some of its properties are stated and proved. These properties will be used to prove Theorems B.7.1 and B.7.2 below.

$$\bar{c}_{nj} \triangleq \frac{n g(x_j)}{2 f_c} \left(\sqrt{1 + \frac{4}{n} \frac{f_c}{g^2(x_j)}} - 1 \right) \ \forall 1 \leq j \leq n \tag{B.3a}$$

B.3 Properties of $\bar{c}_{nj}$

$\bar{c}_{nj}$ has following properties:

$$(P1)\ \frac{1}{\bar{c}_{nj}} - \frac{\bar{c}_{nj} f_c}{n} = g(x_j) \tag{B.4a}$$

$$(P2)\ \bar{c}_{nj} = \frac{1}{g(x_j)}\left(1 + \mathcal{O}\left(\frac{1}{ng^2(x_j)}\right)\right) \quad \text{for} \quad ng^2(x_j) > f_c \tag{B.4b}$$

$$(P3)\ \bar{c}_{nj}^2 = \frac{n}{f_c}(1 - \bar{c}_{nj} g(x_j)) \tag{B.4c}$$

$$(P4)\ \sqrt{\frac{3/2 - \sqrt{5}/2}{f_c}} \leq |\bar{c}_{nj}| \leq \sqrt{\frac{n}{f_c}} \tag{B.4d}$$

$$(P5)\ 0 \leq 1 - \bar{c}_{nj} g(x_j) \leq 1 \tag{B.4e}$$

$$(P6)\ 1 - \frac{1}{n}\sum_{j=1}^{n} \bar{c}_{nj} g(x_j) < \mathcal{O}_{a.s.}\left(\frac{1}{\sqrt{n}}\right)\ if\ g(x) > 0 \forall x \tag{B.4f}$$

$$(P7)\ \frac{1}{n}\sum_{j \neq i}(s_{ij}\bar{c}_{nj}) = \frac{1}{n}\sum_{j} s_{ij}\bar{c}_{nj} - \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$$
$$= g(x_i) + \epsilon_{ni} \overset{a.s.}{\rightarrow} g(x_i) \text{ simultaneously } \forall i\ if\ g(x) > 0 \forall x \tag{B.4g}$$

$$(P8)\ \bar{c}_{\infty j} \triangleq \lim_{n \to \infty} \bar{c}_{nj} \geq \bar{c}_{nj}\ \forall n \tag{B.4h}$$

B.4 Proofs for properties of $\bar{c}_{nj}$

(P1) can be proved by direct substitution of $\bar{c}_{nj}$ into left hand side (LHS). (P2) can be derived through binomial expansion of $\bar{c}_{nj}$. (P3) can again be proved by substituting $\bar{c}_{nj}$ and showing LHS=RHS. (P4) and (P5) can be proved by using the fact that both $\bar{c}_{nj}^2$ and $\bar{c}_{nj} g(x_j)$ are monotonic in $g^2(x_j)$ since $\frac{d\bar{c}_{nj}^2}{dg^2(x_j)} < 0$ and $\frac{d\bar{c}_{nj} g(x_j)}{dg^2(x_j)} > 0$. Therefore, the minimum and maximum values of $|c_j|$ and $c_j g(x_j)$, can be found in by plugging in the

minimum and maximum values of $g^2(x_j)$ (note $0 \leq g^2(x_j) \leq f_c$, from Thm B.1.1).

(P6) is proved by using Kolmogorov's sufficient criterion (102) for almost sure convergence of sample mean. Clearly, from (P5) $E(\bar{c}_{nj}^2 g^2(x_j)) < \infty$ which establish almost sure convergence. Now, let $\beta \triangleq \frac{1}{n}\sum \bar{c}_{nj} g(x_j)$. Then multiplying each side of n equations in (P1) by $\frac{1}{g(x_j)}$, respectively, adding them and the normalizing the sum by $\frac{1}{n}$ gives:

$$\frac{1}{n}\sum \frac{1}{\bar{c}_{nj} g(x_j)} = 1 + \frac{1}{n}\sum \frac{\bar{c}_{nj} f_c}{n g(x_j)} \tag{B.5a}$$

$$\Rightarrow \frac{1}{\beta} \leq 1 + b_n \tag{B.5b}$$

$$\Rightarrow \beta \geq \frac{1}{1+b_n} \tag{B.5c}$$

Above $b_n \triangleq \sum_j \frac{f_c \bar{c}_{nj}}{n^2 g(x_j)}$. To go from (B.5)a to (B.5)b, the result $\frac{1}{n}\sum \frac{1}{\bar{c}_{nj} g(x_j)} \geq \frac{n}{\sum \bar{c}_{nj} g(x_j)} = \frac{1}{\beta}$ (Arithmetic Mean $\geq$ Harmonic Mean) is used. Now it can be shown that $b_n \leq \mathcal{O}_{a.s}\left(\frac{1}{\sqrt{n}}\right)$, as following:

$$b_n = \sum_i \frac{f_c \bar{c}_{ni}}{n^2 g(x_i)} \tag{B.6a}$$

$$\leq \frac{\sqrt{f_c}}{n\sqrt{n}} \sum_i \frac{1}{g(x_i)} \tag{B.6b}$$

$$\overset{a.s}{\to} \sqrt{\frac{f_c}{n}} E\left(\frac{1}{g(x_i)}\right) \tag{B.6c}$$

$$= \mathcal{O}_{a.s}\left(\frac{1}{\sqrt{n}}\right) \tag{B.6d}$$

To go from (B.6)a to (B.6)b (P4) and $g(x) > 0$ are used. To go from (B.6)c to (B.6)d, $E\left(\frac{1}{g(x_i)}\right) = \int g(x_i) dx_i$ is used, which has to be bounded as $g^2(x)$ is pdf and bandlimited (due to Plancherel). Finally the fact that the sample mean of positive numbers, if converges,

converges almost surely gives (B.6)d. Combining (B.6)d and (B.5)c gives:

$$\beta \geq 1 - \mathcal{O}_{a.s}\left(\frac{1}{\sqrt{n}}\right) \tag{B.7}$$

substituting β in LHS of (P6) proves it.

To prove (P7) Kolmogorov's sufficient criterion (102) is first used to establish the almost sure convergence of each equation separately. Due to Kolmogrov's sufficient criterion:

$$\frac{1}{n}\sum_{j\neq i} s_{ij}\bar{c}_{nj} = \frac{1}{n}\sum_{j} s_{ij}\bar{c}_{nj} - \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) \tag{B.8a}$$

$$\xrightarrow{a.s.} E_j(s_{ij}\bar{c}_{nj}) \;\; if \; E_j(\bar{c}_{nj}^2 s_{ij}^2) < \infty \tag{B.8b}$$

Thus, now $E_j(\bar{c}_{nj}s_{ij})$ and $E_j(\bar{c}_{nj}^2 s_{ij}^2)$ are computed as follows:

$$\begin{aligned}
&|E_j(\bar{c}_{nj}s_{ij}) - g(x_i)| \\
&= \left|\int \bar{c}_{nj}s_{ij}g^2(x_j)\mathrm{d}x_j - g(x_i)\right| && \text{(B.9a)}\\
&= \left|\int_{ng^2(x)\geq f_c} s_{ij}g(x_j) + \mathcal{O}\left(\frac{1}{n}\right)\frac{s_{ij}}{g(x_j)}\mathrm{d}x_j \right. \\
&\quad \left. + \int_{ng^2(x)<f_c} \bar{c}_{nj}s_{ij}g^2(x_j)\mathrm{d}x_j - g(x_i)\right| && \text{(B.9b)}\\
&= \left|\int s_{ij}g(x_j)\mathrm{d}x_j - \int_{ng^2(x)<f_c}(1-\bar{c}_{nj}g(x_j))s_{ij}g(x_j)\mathrm{d}x_j \right. \\
&\quad \left. + \int_{ng^2(x)\geq f_c} \mathcal{O}\left(\frac{1}{n}\right)\frac{s_{ij}}{g(x_j)}\mathrm{d}x_j - g(x_i)\right| && \text{(B.9c)}\\
&\leq \int_{ng^2(x)<f_c} |s_{ij}g(x_j)|\mathrm{d}x_j + \mathcal{O}\left(\frac{1}{n}\right)\int_{ng^2(x)\geq f_c}\left|\frac{s_{ij}}{g(x_j)}\right|\mathrm{d}x_j && \text{(B.9d)}
\end{aligned}$$

To go from (B.9)c to (B.9)d, the facts that $\int s_{ij} g(x_j) \mathrm{d}x_j = g(x_i)$ for any $g \in \mathbb{V}(\omega_c)$ and (P5) are used. Now define

$$\varepsilon_n(x_i) \triangleq \mathcal{O}\left(\tfrac{1}{n}\right) \int_{ng^2(x)\geq f_c} \left|\tfrac{s_{ij}}{g(x_j)}\right| \mathrm{d}x_j + \int_{ng^2(x)<f_c} |s_{ij} g(x_j)| dx_j.$$

Then it is shown,

$$|E_j(\bar{c}_{nj} s_{ij}) - g(x_i)| \leq \varepsilon_n(x_i) \to 0 \quad \text{uniformly } if\ g(x) > 0 \tag{B.10a}$$

by first noting that

$$\int_{ng^2(x)\geq f_c} \left|\frac{s_{ij}}{g(x_j)}\right| \mathrm{d}x_j \leq \sqrt{\frac{n}{f_c}} \int_{ng^2(x)\geq f_c} |s_{ij}| \, \mathrm{d}x_j,$$

and that the length of limit of integration has to be less than $\frac{n}{f_c}$ as $g^2(x)$ has to integrate to 1. This makes $\int_{ng^2(x)\geq f_c} |s_{ij}| \, \mathrm{d}x_j \leq \mathcal{O}(\log(n))$ and hence

$$\mathcal{O}\left(\frac{1}{n}\right) \int_{ng^2(x)\geq f_c} \left|\frac{s_{ij}}{g(x_j)}\right| \mathrm{d}x_j \leq \mathcal{O}\left(\frac{\log(n)}{\sqrt{n}}\right) \to 0 \text{ uniformly.}$$

Then, $\int_{ng^2(x)<f_c} |s_{ij} g(x_j)| dx_j < f_c \int_{ng^2(x)<f_c} g(x_j) dx_j \quad if\ g(x) > 0$ is also shown to go to zero uniformly, by first considering

$$\zeta_n(x_j) \triangleq \begin{cases} g(x_j) & \text{if } g^2(x_j) \geq \frac{f_c}{n} \\ 0 & o.w. \end{cases} \tag{B.11}$$

The sequence $\zeta_n(x_j)$ is non-decreasing under the condition $g^2(x) > 0$ & $g^2(x) \in \mathbb{U}(\omega_c)$, i.e $\zeta_{n+1}(x_j) \geq \zeta_n(x_j)\ \forall\ x_j$, and the $\lim_{n\to\infty} \zeta_n(x_j) = g(x_j)$. Therefore, by the monotone

convergence theorem, $\lim_{n\to\infty}\int\zeta_n(x_j)dx_j = \int_{-\infty}^{\infty} g(x_j)dx_j$. This limit converges due to Plancherel. Now, by definition of $\zeta_n(x_j)$,

$$\lim_{n\to\infty}\int_{ng^2(x)<f_c}|s_{ij}g(x_j)|dx_j \leq$$
$$f_c\int_{-\infty}^{\infty}g(x_j)dx_j - \lim_{n\to\infty}f_c\int\zeta_n(x_j)dx_j \to 0 \text{ uniformly.} \tag{B.12a}$$

Therefore $\varepsilon_n(x_i)\to 0$ uniformly $\forall x_i$ which is equivalent to saying $\max_x\varepsilon(x)\to 0$. A weaker but more informative proof for going to step (B.9)e to (B.9)d can be obtained by assuming a tail behaviour of $\frac{1}{|x|^r}$ for $g^2(x)$ and showing the step holds for all $r>1$, this gives $\varepsilon_n(x_i)=\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\ \forall x_i$. Now it is shown:

$$E_j(\bar{c}_{nj}^2 s_{ij}^2) = \int\bar{c}_{nj}^2 s_{ij}^2 g^2(x_j)dx_j \tag{B.13a}$$
$$\leq \int s_{ij}^2 dx_j = f_c < \infty\forall x_i \tag{B.13b}$$

To go from (B.13)a to (B.13)b, (P5) and the equality $\int s_{ij}^2 dx_j = f_c$ are invoked. Finally, substituting (B.9)f and (B.13)b into (B.8)b proves that each equation go to zero almost surely but separately. More precisely, until now only it has been shown that there exists sets of events $E_1, E_2, \cdots, E_n$ where each set $E_i \triangleq \{s : \lim_{n\to\infty}\rho_{ni}(\bar{c}(s)) = 0\}$ and $P(E_i)=1$. However to establish simultaneity of convergence it is further needed to be shown that $P(\cap_i^{\infty}E_i)=1$.

For this, the almost sure convergence of the following L^2 norm:

$$\int \left(\frac{1}{n} \sum \bar{c}_{nj} s(x - x_j) - g(x) \right)^2 \mathrm{d}x \overset{a.s.}{\to} 0 \quad \text{if } g(x) > 0 \tag{B.14}$$

is established in next section. This implies that $\frac{1}{n} \sum \bar{c}_{nj} s(x - x_j) \overset{a.s.}{\to} g(x)$ uniformly due to band-limitedness of functions (103). This in turn implies that eqns $\frac{1}{n} \sum_j \bar{c}_{nj} s(x_i - x_j) \overset{a.s.}{\to} g(x_i)$ simultaneously for all x_i and hence prove (P7).

(P8) can be proved easily by showing that $\frac{d\bar{c}_{nj}}{dn} > 0 \; \forall n$.

B.5 Proof for (B.14)

To establish convergence of L^2 norm consider:

$$\begin{aligned}
&\int \left(\frac{1}{n} \sum \bar{c}_{nj} s(x - x_j) - g(x) \right)^2 \mathrm{d}x \\
&= \int \frac{1}{n^2} \sum_{ij} \bar{c}_{ni} \bar{c}_{nj} s(x - x_i) s(x - x_j) + g^2(x) \\
&- \frac{2}{n} \sum_j \bar{c}_{nj} s(x - x_j) g(x) \mathrm{d}x && \text{(B.15a)} \\
&= \frac{1}{n^2} \sum_{ij} \bar{c}_{ni} \bar{c}_{nj} s_{ij} + 1 - \frac{2}{n} \sum_j \bar{c}_{nj} g(x_j) && \text{(B.15b)} \\
&= \frac{1}{n^2} \sum_{i \neq j} \bar{c}_{ni} \bar{c}_{nj} s_{ij} + \frac{1}{n^2} \sum_i s_{ii} \bar{c}_{ni}^2 + 1 - \frac{2}{n} \sum_j \bar{c}_{nj} g(x_j) && \text{(B.15c)} \\
&= \frac{1}{n^2} \sum_{i \neq j} \bar{c}_{ni} \bar{c}_{nj} s_{ij} + \frac{1}{n} \sum_i (1 - \bar{c}_{ni} g(x_i)) + 1 - \frac{2}{n} \sum_j \bar{c}_{nj} g(x_j) && \text{(B.15d)} \\
&\overset{a.s.}{\to} E(\bar{c}_{ni} \bar{c}_{nj} s_{ij}) - 1 && \text{(B.15e)}
\end{aligned}$$

To go from (B.15)c to (B.15)d to (B.15)d P3 and P6. For going to (B.15)e the almost sure convergence proof is established in next sectionB.6.

Now, $E(\bar{c}_{ni}\bar{c}_{nj}s_{ij})$ is calculated as:

$$E(\bar{c}_{ni}\bar{c}_{nj}s_{ij}) = \int \bar{c}_{ni}\bar{c}_{nj}s_{ij}g^2(x_i)g^2(x_j)\mathrm{d}x_i\mathrm{d}x_j \tag{B.16a}$$

$$= \int \bar{c}_{ni}g^2(x_i)(g(x_i)+\varepsilon_n(x_i))\mathrm{d}x_i \tag{B.16b}$$

$$= \int \bar{c}_{ni}g^3(x_i)\mathrm{d}x_i + \int \bar{c}_{ni}g^2(x_i)\varepsilon_n(x_i)\mathrm{d}x_i \tag{B.16c}$$

$$= 1 + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) + \max_{x_i}(\varepsilon_n(x_i))\int |g(x_i)|\mathrm{d}x_i \tag{B.16d}$$

$$\to 1 \qquad \text{if } g(x) > 0 \tag{B.16e}$$

To go from (B.16)a to (B.16)b (B.10) is used. To go from (B.16)c to (B.16)d (P6) and (P5) are used. To go from (B.16)d to (B.16)e uniform convergence of $\varepsilon_n(x)$ and $\int g(x) < \infty$ (due to Plancheral) are used. Now, combining (B.16) e and (B.15)e establishes (B.14) and subsequently simultaneous convergence, in almost sure sense.

B.6 Proof for almost sure convergence of $\frac{1}{n^2}\sum_{i\neq j}\bar{c}_{ni}\bar{c}_{nj}s_{ij}$

Let $S_n \triangleq \frac{1}{n^2}\sum_{i\neq j}\bar{c}_{ni}\bar{c}_{nj}s_{ij}$, then:

$$
\begin{aligned}
Var\left(S_n\right) &= \frac{4n(n-1)(n-2)}{n^4}E\left(\bar{c}_{ni}\bar{c}_{nj}^2\bar{c}_{nm}s_{ij}s_{jm}\right) \\
&+ \frac{2n(n-1)}{n^4}E\left(\bar{c}_{ni}^2\bar{c}_{nj}^2s_{ij}^2\right) && \text{(B.17a)} \\
&- \frac{2n(n-2)(2n-3)}{n^4}E\left(\bar{c}_{ni}\bar{c}_{nj}\bar{c}_{nl}\bar{c}_{nm}s_{ij}s_{lm}\right) \\
&\leq \frac{4n(n-1)(n-2)f_c}{n^4}\left(\int g(x_i)\mathrm{d}x_i\right)^2 \\
&+ \frac{2n(n-1)}{f_cn^3}E\left(\bar{c}_{ni}^2(1-\bar{c}_{nj}g(x_j))s_{ij}^2\right) \\
&+ \frac{2n(n-2)(2n-3)}{n^4}E\left(\bar{c}_{ni}\bar{c}_{nj}|s_{ij}|\right)^2 && \text{(B.17b)} \\
&= \mathcal{O}\left(\frac{1}{n}\right) && \text{(B.17c)}
\end{aligned}
$$

To go from (B.17)a to (B.17)b $\int|s_{ij}s_{jm}| < f_c$ (Cauchy-Schwartz inequality), P5, P3 are used. To go from (B.17)b to (B.17)c $\int g(x) < \infty$ (due to Plancheral), P5 and $\int|s_{ij}g(x_i)| < \sqrt{f_c}$ (Cauchy-Schwartz inequality) are used.

Now, by Chebyshev inequality $\Pr(|S_{n^2}-\mu| > \epsilon) < \mathcal{O}\left(\frac{1}{n^2}\right)$, here $\mu = \lim_{n\to\infty}E(S_n)$. Hence, $\sum_{n=1}^{\infty}\Pr(|S_{n^2}-\mu| > \epsilon) < \infty$, therefore by Borel-Cantelli lemma, $S_{n^2} \overset{a.s.}{\to} \mu$. Now to show $S_n \overset{a.s.}{\to} \mu$, divide S_n into two parts $A_n \triangleq \frac{1}{n^2}\sum_{i\neq j}\bar{c}_{ni}\bar{c}_{nj}s_{ij}I(s_{ij})$ where $I(s_{ij})$ is indicator function which if 1 is $s_{ij} \geq 0$ and 0 otherwise (note that $\bar{c}_{ni} > 0\ \forall i$ due to the

assumption $g(x) > 0$), and $B_n \triangleq S_n - A_n$. Now,

$$\begin{aligned}
Var\left(A_n\right) &< \frac{4n(n-1)(n-2)}{n^4}E\left(\bar{c}_{ni}\bar{c}_{nj}^2\bar{c}_{nm}|s_{ij}||s_{jm}|\right) \\
&+ \frac{2n(n-1)}{n^4}E\left(\bar{c}_{ni}^2\bar{c}_{nj}^2 s_{ij}^2\right) && \text{(B.18a)} \\
&- \frac{2n(n-2)(2n-3)}{n^4}E\left(\bar{c}_{ni}\bar{c}_{nj}\bar{c}_{nl}\bar{c}_{nm}|s_{ij}||s_{lm}|\right) \\
&\leq \frac{4n(n-1)(n-2)f_c}{n^4}\left(\int g(x_i)\mathrm{d}x_i\right)^2 \\
&+ \frac{2n(n-1)}{f_c n^3}E\left(\bar{c}_{ni}^2(1-\bar{c}_{nj}g(x_j))|s_{ij}|^2\right) \\
&+ \frac{2n(n-2)(2n-3)}{n^4}E\left(\bar{c}_{ni}\bar{c}_{nj}|s_{ij}|\right)^2 && \text{(B.18b)} \\
&= \mathcal{O}\left(\frac{1}{n}\right) && \text{(B.18c)}
\end{aligned}$$

To go from (B.18)a to (B.18)b $\int |s_{ij}s_{jm}| < f_c$ (Cauchy-Schwartz inequality), P5, P3 are used. To go from (B.17)b to (B.17)c $\int g(x) < \infty$ due to plancheral, P5 and $\int |s_{ij}g(x_i)| < \sqrt{f_c}$ (Cauchy-Schwartz inequality) are used. Therefore, again by Chebyshev inequality and Borel-Cantelli lemma (104) $A_{n^2} \overset{a.s.}{\to} \lim_{n\to\infty} E(A_n)$. Now, consider integer k such that $k^2 \leq n \leq (k+1)^2$, as n^2A_n is monotonically increasing (by definition) this implies:

$$\frac{k^4}{(k+1)^2}A_{k^2} \leq A_n \leq \frac{(k+1)^4}{k^4}A_{(k+1)^2} \tag{B.19a}$$

$$\overset{a.s.}{\to} \lim_{n\to\infty} E(A_n) \leq A_n \leq \lim_{n\to\infty} E(A_n) \tag{B.19b}$$

Now by sandwich theorem $A_n \overset{a.s.}{\to} \lim_{n\to\infty} E(A_n)$, similarly it can be shown that $B_n \overset{a.s.}{\to} \lim_{n\to\infty} E(B_n)$ and hence $S_n \overset{a.s.}{\to} \lim_{n\to\infty} E(S_n)$. Hence proved.

Now, Theorems B.7.1 and B.7.2 are proven.

B.7 Proof for consistency of the BLML estimator

Theorem B.7.1 *Suppose that the observations, x_i for $i = 1, ..., n$ are i.i.d. and distributed as $x_i \sim g^2(x) \in \mathbb{U}(\omega_c)$. Then, $\bar{c}_{\infty i} \triangleq \lim_{n\to\infty} \frac{ng(x_i)}{2f_c}\left(\sqrt{1 + \frac{4}{n}\frac{f_c}{g^2(x_i)}} - 1\right)$ is a solution to $\boldsymbol{\rho_n}(\mathbf{c}) = \mathbf{0}$ in the limit as $n \to \infty$.*

Proof: To prove this theorem, we establish that any equation $\rho_{ni}(\bar{\mathbf{c}}_\mathbf{n})$, indexed by i goes to 0 almost surely as $n \to \infty$ as follows:

$$\rho_{ni}(\bar{\mathbf{c}}_\mathbf{n}) = \frac{1}{n}\sum_{j\neq i} s_{ij}\bar{c}_{nj} + \frac{\bar{c}_{ni}f_c}{n} - \frac{1}{\bar{c}_{ni}} \;\forall i = 1, \cdots, n \tag{B.20a}$$

$$\stackrel{a.s.}{\to} g(x_i) - g(x_i) = 0 \;\forall i = 1, \cdots, n \tag{B.20b}$$

In moving from (B.20)a to (B.20)b (P1) and (P7) are used. (B.20)b, show that each of the $\rho_{ni}(\bar{c}_n)\;\forall i$ goes to 0 in probability. Therefore,

$$\lim_{n\to\infty} \rho_{ni}(\bar{\mathbf{c}}_\mathbf{n}) = 0 \;\forall i = 1, \cdots, n \tag{B.21}$$

This proves Theorem B.7.1. Note that one may naively say that $\lim_{n\to\infty} \bar{c}_{ni} = \frac{1}{g(x_i)} \;\forall\, i = 1, \cdots, n$ (see (P2)). However, this is not true because even for large n there is a finite probability of getting at least one $g(x_i)$ which is so small such that $\frac{1}{ng^2(x_i)}$ may be finite, and hence $\lim_{n\to\infty} \bar{c}_{ni}$ cannot be calculated the usual way. Therefore, it is wise to write down

$\bar{c}_{\infty i} \triangleq \lim_{n\to\infty} \bar{c}_{ni}$ as a solution to (A.14), instead of $\frac{1}{g(x_i)}$.

Theorem B.7.2 *Suppose that the observations, x_i for $i = 1, ..., n$ are i.i.d. and distributed as $x_i \sim f(x) \in \mathbb{U}(\omega_c)$ and $f(x) > 0\ \forall x$. Let, $f_\infty(x) \triangleq \lim_{n\to\infty} \left(\frac{1}{n}\sum_{i=1}^{n} \bar{c}_{\infty i} \frac{\sin(\pi f_c(x-x_i))}{\pi(x-x_i)}\right)^2$, then $\int \left(f(x) - f_\infty(x)\right)^2 \mathrm{d}x = 0$.*

Proof: Let $\hat{g}_\infty(x) \triangleq \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \bar{c}_{\infty i} s(x - x_i)$ here $s(x - x_i) \triangleq \frac{\sin(\pi f_c(x-x_i))}{\pi(x-x_i)}$. Therefore the ISE is:

$$
\begin{aligned}
ISE &\triangleq \int \left(\hat{g}^2_\infty(x) - g^2(x)\right)^2 \mathrm{d}x && \text{(B.22a)}\\
&= \int \left(\hat{g}(x) - \hat{g}_\infty(x)\right)^2 \left(\hat{g}_\infty(x) + g(x)\right)^2 \mathrm{d}x && \text{(B.22b)}\\
&\leq 4f_c \int (\hat{g}_\infty(x) - g(x))^2 \mathrm{d}x && \text{(B.22c)}\\
&= 4f_c \int \lim_{n\to\infty} \left(\frac{1}{n}\sum \bar{c}_{nj} s(x - x_j) - g(x)\right)^2 \mathrm{d}x && \text{(B.22d)}\\
&\leq 4f_c \liminf_{n\to\infty} \int \left(\frac{1}{n}\sum \bar{c}_{nj} s(x - x_j) - g(x)\right)^2 \mathrm{d}x && \text{(B.22e)}\\
&\overset{a.s.}{\to} 0 && \text{(B.22f)}
\end{aligned}
$$

To go from (B.22)b to (B.22)c, the inequality $(g(x) + \hat{g}(x))^2 \leq 4f_c$ if $\hat{g}, g \in \mathbb{V}$ is used (see Theorem B.1.1). To go from (B.22)c to (B.22)d, $\hat{g}_\infty(x)$ is expanded. To go from (B.22)d to (B.22)e, Fatou's lemma (105) is invoked as the function inside the integral is non-negative and measurable. In particular, due to (P6), $\phi_n(x) = \frac{1}{n}\sum \bar{c}_{nj} s(x - x_j) - g(x) \overset{a.s.}{\to} E(\bar{c}_{nj} s(x - x_j)) - g(x) = 0$, which establishes the point-wise convergence of $\phi^2_n(x)$ to 0. Hence, lim can be safely replaced by lim inf and Fatou's lemma (106) can be applied. To go from (B.22)e to (B.22)f, (B.14) is used.

Hence proved.

Theorem B.7.3 *Suppose that the observations, x_i for $i = 1, ..., n$ are i.i.d. and distributed as $x_i \sim f(x) \in \mathbb{U}(\omega_c)$. Then, the KL-divergence between $f(x)$ and $f_\infty(x)$ is zero and hence $\bar{\mathbf{c}}_\infty$ is the solution of (7.2) in the limit $n \to \infty$. Therefore, the BLML estimator $\hat{f}(x) = f_\infty(x) = f(x)$ in probability.*

Proof: Consider $\{x_1, \cdots, x_n\}$ to be a member of typical set (69). Then the KL-divergence between $f(x)$ and $f_\infty(x)$ can be bounded as:

$$0 \leq E\left(\log\left(\frac{f(x)}{f_\infty(x)}\right)\right) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{g^2(x_i)}{g_\infty^2(x_i)}\right) \tag{B.23a}$$

$$\leq \lim_{n\to\infty} \frac{2}{n}\sum_{i=1}^{n} \log\left(|g(x_i)\bar{c}_{\infty i}|\right) \tag{B.23b}$$

$$\leq 0 \tag{B.23c}$$

To go from (B.23) a to (B.23) b, definition of g_∞ and P7 is used. To go from (B.23) b to (B.23)c, (P5) is used.

Therefore, the KL divergence between $\hat{f}_\infty(x)$ and the true pdf is 0 and hence $\hat{f}_\infty(x)$ should also maximizes the likelihood function. Finally, $\hat{\mathbf{c}} = \bar{\mathbf{c}}_\infty$ or $\hat{\mathbf{c}} = -\bar{\mathbf{c}}_\infty$. The negative solution can be safely ignored by limiting only to positive solutions. Hence Proved.

Theorem B.7.4 *If* $g^2(x) = f(x) \in \mathbb{U}(\omega_c)$ *such that* $f(x) > 0 \quad \forall\ x \in \mathbb{R}$, *then* $g(x) > 0 \quad \forall\ x \in \mathbb{R}$, *and the asymptotic solution of* (7.2) *lies in the orthant with indicator vector* $c_{0i} = 1\ \forall i = 1, \cdots, n$.

Proof: $g \in \mathbb{V}(\omega_c)$ as $g^2 \in \mathbb{U}(\omega_c)$. Therefore $g(x)$ is band-limited and hence continuous. Now, assume that $\exists\ x_1, x_2 \in \mathbb{R}$ such that $g(x_1) > 0$ and $g(x_2) < 0$. Due to continuity of g this would imply that $\exists\, x_3,\ x_1 < x_3 < x_2$ such that $g(x_3) = f(x_3) = 0$. This is a contradiction as $f(x) > 0\ \forall x \in \mathbb{R}$. Therefore, either $g(x) < 0\ \forall\ x \in \mathbb{R}$ or equivalently $g(x) > 0\ \forall\ x \in \mathbb{R}$. Now, by Theorems B.7.1 and B.7.3, $c_{0i} = sign(\hat{c}_i) = sign(c_{\infty i}) = sign(g(x_i)) = 1\ \forall i = 1 \cdots n$ asymptotically. Hence proved.

Appendix C

Generalization of BLML estimator to joint pdfs

BLML estimator for joints pdfs can be found in a very similar way as it is found for one dimensional pdfs. The only change occurs while defining (A.5), where one needs to define multidimensional $b_i's$ such that

$$b_i(\boldsymbol{\omega}) \triangleq \left\{ \begin{array}{cc} e^{-j\boldsymbol{\omega}^T \mathbf{x}_i} & \forall\ |\boldsymbol{\omega}| \leq \frac{\boldsymbol{\omega}_c}{2} \\ 0 & o.w. \end{array} \right\}, \tag{C.1}$$

inverse Fourier transform of which gives a multidimensional $\text{sinc}_{\mathbf{f_c}}(\cdot)$ function.

Appendix D

`BLMLQuick` Algorithm

Consider a function $\bar{f}(x)$ such that:

$$\bar{f}(x) = f_s \int_{x-\frac{0.5}{f_s}}^{x+\frac{0.5}{f_s}} f(\tau)d\tau \tag{D.1}$$

where $f \in \mathbb{U}(\omega_c)$ and $f_s > 2f_c$ is the sampling frequency. It is easy to verify that $\bar{f}(x)$ is also a pdf and $\bar{f} \in \mathbb{U}(\omega_c)$. Now consider samples $\bar{f}[p] = \bar{f}(p/f_s)$, clearly these samples are related to $f(x)$ as:

$$\bar{f}[p] = \int_{\frac{p-0.5}{f_s}}^{\frac{p+0.5}{f_s}} f(x)dx \tag{D.2}$$

Further consider $\bar{x}_i$'s computed by binning from x_i's, n i.i.d observations of r.v. $x \sim f(x)$, as:

$$\bar{x}_i = f_s \lfloor \frac{x_i}{f_s} - 0.5 \rfloor \tag{D.3}$$

where $\lfloor\rfloor$ is the greatest integer function. Note that $\bar{x}_i$ are the i.i.d. observations from $\tilde{f}(x) = \sum_p \bar{f}[p]\delta\left(x - \frac{p}{f_s}\right)$. Now since $f_s > 2f_c$, the BLML estimate for $\tilde{f}(x)$ should converge

to $\bar{f}(x)$ due to Nyquist's Sampling Theorem (107). This estimator is called `BLMLQuick`. Assuming that the rate of convergence for BLML is $\mathcal{O}(n^{-1})$, then if f_s is chosen such that $||f - \bar{f}||_2^2 = \mathcal{O}(n^{-1})$, the `BLMLQuick` will also converge with $\mathcal{O}(n^{-1})$. This happens at $f_s = f_c n^{0.25} > f_c$ also $f_s > 2f_c$ if $n > 16$.

D.1 Implementation and computational complexity

Before implementing `BLMLQuick` and computing its computational complexity, the following theorem is first stated and proved.

Theorem D.1.1 *Consider n i.i.d observations $\{x_i\}_{i=1}^n$ of random variable x with pdf having $\frac{1}{|x|^r}$ tail. Then*

$$\Pr\left(\min(\{x_i\}_{i=1}^n) < -\left(\frac{n}{(r-1)\epsilon}\right)^{\frac{1}{r-1}}\right) \simeq 1 - e^{-\epsilon} \simeq \epsilon \tag{D.4}$$

for large n.

Proof: For n i.i.d observations $\{x_i\}_{i=1}^n$ of random variable x with cumulative distribution function $F(x)$, it is well known that :

$$\Pr(\min(\{x_i\}_{i=1}^n) < x) = 1 - (1 - F(x))^n \tag{D.5a}$$

$$\simeq 1 - e^{-nF(x)} \ \forall F(x) < 0.5 \tag{D.5b}$$

$$\simeq 1 - e^{-\frac{n}{(r-1)|x|^{r-1}}} \ \forall F(x) < 0.5 \tag{D.5c}$$

substituting $x = -\left(\frac{n}{(r-1)\epsilon}\right)^{\frac{1}{r-1}}$ above proves the result.

Finally, due to duplicity in $\bar{x}_i$ $i = 1, \ldots, n$, they can be written concisely as $[\bar{x}_b, n_b]$, $b = 1, \ldots, B$ where $\bar{x}_b$ are unique values in $\bar{x}_i$ and n_b is duplicity count of $\bar{x}_b$. Now it can be observed that $B \leq (\max(x_i) - \min(x_i))f_s \leq \mathcal{O}_p(n^{\frac{1}{r-1}})f_s$, if the true pdf has tail that decreases as $\frac{1}{|x|^r}$ (Theorem D.1.1).

Now the `BLMLQuick` is implemented using following steps:

- Compute $\{\bar{x}_b, n_b\}_{b=1}^B$ from $\{x_i\}_{i=1}^n$. Computational complexity of $\mathcal{O}(n)$.
- Sort $\{\bar{x}_b, n_b\}_{b=1}^B$ and construct $\mathbf{S} : s_{ab} = s(\bar{x}_a - \bar{x}_b)\ \forall\ a, b = 1, \ldots, B$ and $\bar{\mathbf{S}} = \mathbf{S} * diag(\{n_b\}_{b=1}^B)$. Note that $\mathbf{S}$ is block-Toeplitz matrix (Toeplitz arrangements of blocks and each block is Toeplitz) (108). Computational complexity of $\mathcal{O}(B^2)$.
- Use convex optimization algorithms to solve $\rho_n(\mathbf{c}) = 0$. Newton's method should take a finite number of iterations to reach a given tolerance ϵ since the cost function is self concordant (54). Therefore, the computational complexity of optimization is same as the computational complexity of one iteration. The complexity of one iteration is the same as the complexity of calculating

$$\left(diag\left(\{1/c_b^2\}_{b=1}^B\right) + \mathbf{S} \times diag\left(\{n_b\}_{b=1}^B\right)\right)^{-1} \tag{D.6a}$$

$$= \left(diag\left(\{1/(c_b^2 n_b)\}_{b=1}^B\right) + \mathbf{S}\right)^{-1} diag\left(\{n_b\}_{b=1}^B\right)^{-1} \tag{D.6b}$$

  As $diag\left(\{1/(c_b^2 n_b)\}_{b=1}^B\right) + \mathbf{S}$ is also block-Toeplitz structure, the Akaike algorithm (108) can be used to evaluate each iteration of Newton's method in $\mathcal{O}(B^2)$.

  Note: Simulations show that $\mathbf{S}$ can be approximated accurately (to machine accuracy)

by a low rank matrix e.g., $R = 20$ for $B = 1000$, therefore the inversion can be performed in $\mathcal{O}(R^2 + RB)$. Further, in some cases one may end up with a large B (e.g. if true pdf has heavy tails) so that storing the Hessian matrix becomes expensive. In such cases, a quasi Newton or gradient descent can be used which compute BLML estimator fairly quickly.

- Evaluate `BLMLQuick` estimate $f(x) = (\frac{1}{n}\sum_{b=1}^{B} n_b c_b s(x - x_b))^2$ at l given points. Computational complexity of $\mathcal{O}(Bl)$.

  The total computational complexity is $\mathcal{O}(n+B^2+lB)$. Substituting $B \leq \mathcal{O}\left(n^{\frac{1}{r-1}}\right) f_s \leq \mathcal{O}\left(f_c n^{\frac{1}{r-1}+0.25}\right)$, gives the total computational complexity
  $\mathcal{O}\left(n + f_c^2 n^{\frac{2}{r-1}+0.5} + f_c l n^{\frac{1}{r-1}+0.25}\right)$.

References

[1] Jake Ormond and Bruce L McNaughton. **Place field expansion after focal MEC inactivations is consistent with loss of Fourier components and path integrator gain reduction**. *Proceedings of the National Academy of Sciences*, **112**(13):4116–4121, 2015. iv, 5

[2] André A Fenton and Robert U Muller. **Place cell discharge is extremely variable during individual passes of the rat through the firing field**. *Proceedings of the National Academy of Sciences*, **95**(6):3182–3187, 1998. iv, 76, 85, 90

[3] Michael S Lewicki. **A review of methods for spike sorting: the detection and classification of neural action potentials**. *Network: Computation in Neural Systems*, **9**(4):R53–R78, 1998. 2

[4] John O'Keefe. **A review of the hippocampal place cells**. *Progress in neurobiology*, **13**(4):419–439, 1979. 2, 76

[5] Michael E Hasselmo. **Grid cell mechanisms and function: contributions of entorhinal persistent spiking and phase resetting**. *Hippocampus*, **18**(12):1213–1229, 2008. 2, 76

[6] Cox DR and Isham V. *Point Processes*. Chapman and Hall/CRC, 2000. 3, 4

[7] Rahul Agarwal and Sridevi V Sarma. **The effects of dbs patterns on basal ganglia activity and thalamic relay**. *Journal of computational neuroscience*, **33**(1):151–167, 2012. 3

[8] Rahul Agarwal and Sridevi V Sarma. **Restoring the basal ganglia in Parkinson's disease to normal via multi-input phase-shifted deep brain stimulation**. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 1539–1542. IEEE, 2010. 3

[9] Donald L Snyder and Michael I Miller. *Random point processes in time and space*. Springer Science & Business Media, 2012. 4

[10] Peter McCullagh. **Generalized linear models**. *European Journal of Operational Research*, **16**(3):285–292, 1984. 4, 8, 78

[11] Wilson Truccolo, Uri T Eden, Matthew R Fellows, John P Donoghue, and Emery N Brown. **A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects**. *Journal of neurophysiology*, **93**(2):1074–1089, 2005. 4, 78

[12] Liam Paninski. **Maximum likelihood estimation of cascade point-process neural encoding models**. *Network: Computation in Neural Systems*, **15**(4):243–262, 2004. 4

[13] Brown EN, Frank LM, Tang D, Quirk MC, and Wilson MA. **A Statistical Paradigm for Neural Spike Train Decoding Applied to Position Prediction from Ensemble Firing Patterns of Rat Hippocampal Place Cells**. *The Journal of Neuroscience*, **18**(18):7411–25, 1998. 5

[14] Barbieri R, Frank L, Nguyen D, Quirk M, Solo V, Wilson M, and Brown EN. **Dynamic analyses of information encoding in neural ensembles**. *Neural Computation*, **16**:277–307, 2004. 5, 91

[15] Sarma S. V., Eden U.T., Cheng M. L., Williams Z., Hu R., Eskandar E. N., and Brown E. N. **Using Point Process Models to Compare Neural Spiking Activity in the Sub-thalamic Nucleus of Parkinson's Patients and a Healthy Primate**. *IEEE TBME*, **57**(6):1297–305, 2010. 5, 77, 78

[16] Sarma, S.V., Cheng, M., Eden, U., Hu, R., Williams, Z., Brown, E.N., and Eskandar, E. **Modeling neural spiking activity in the sub-thalamic nucleus of Parkinson's patients and a healthy primate**. In *Decision and Control, CDC . 47th IEEE Conference on*, pages 2012 – 2017, 2008. 5

[17] Rahul Agarwal, Sridevi V Sarma, Nitish V Thakor, Marc H Schieber, and Steve Massaquoi. **Sensorimotor Gaussian Fields Integrate Visual and Motor Information in Premotor Neurons**. *J Neurosci*, **35**(25):9508–9525, 2015. 5

[18] Rahul Agarwal and Sridevi V Sarma. **An analytical study of relay neuron's reliability: Dependence on input and model parameters**. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 2426–2429. IEEE, 2011. 5

[19] Rahul Agarwal and Sridevi V Sarma. **Performance limitations of relay neurons**. *PLoS Comput. Biol*, **8**(8):e1002626, 2012. 5

[20] Rahul Agarwal, Sabato Santaniello, and Sridevi V Sarma. **Generalizing performance limitations of relay neurons: Application to Parkinson's disease**. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 6573–6576. IEEE, 2014. 5

[21] Kloosterman F, Layton S, Chen Z, and Wilson MA. **Bayesian decoding using unsorted spikes in the rat hippocampus**. *J. Neurophysiol*, **111**(1):217–27, 2014. 5, 79, 82, 91

[22] John L Kubie and Steven E Fox. **Do the spatial frequencies of grid cells mold the firing fields of place cells?** *Proceedings of the National Academy of Sciences*, **112**(13):3860–3861, 2015. 5, 6, 74, 89

[23] Hanne Stensola, Tor Stensola, Trygve Solstad, Kristian Frøland, May-Britt Moser, and Edvard I Moser. **The entorhinal grid map is discretized**. *Nature*, **492**(7427):72–78, 2012. 5

[24] Bernard W Silverman. *Density estimation for statistics and data analysis*, **26**. CRC press, 1986. 7

[25] Karl Pearson. **Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material**. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, **186**:343–414, 01 1895. 8

[26] Bernard W Silverman. **On the estimation of a probability density function by the maximum penalized likelihood method**. *The Annals of Statistics*, pages 795–810, 1982. 9, 23

[27] Robert Tibshirani. **Regression shrinkage and selection via the lasso**. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 9

[28] Joseph Lee Rodgers and W. Alan Nicewander. **Thirteen Ways to Look at the Correlation Coefficient**. *The American Statistician*, **42**(1):59–66, 1988. 11, 27, 29

[29] Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. **Measuring and testing dependence by correlation of distances**. *The Annals of Statistics*, **35**(6):2769–2794, 2007. 11, 27, 28, 29

[30] Gábor J. Székely and Maria L. Rizzo. **Brownian distance covariance**. *Ann. Appl. Stat.*, **3**(4):1236–1265, 12 2009. 11, 27, 29

[31] C W Granger, E Maasoumi, and J Racine. **A Dependence Metric for Possibly Nonlinear Processes**. *Journal of Time Series Analysis*, **25**(5), 2004. 11, 25, 55, 56, 71

[32] C E Shannon. **A Mathematical Theory of Communication**. *Bell Syst. Tech. J.*, **27**:379–423, 623–656, 1948. 11, 27

[33] Greenstied Charles M. and Snell J. Laurie. *Introduction to Probability*. American Mathematical Society, 2003. 12

[34] Farhad Mehran. **Variance of the MVUE for the lognormal mean**. *Journal of the American Statistical Association*, **68**(343):726–727, 1973. 13

[35] Calyampudi Radhakrishna Rao. **Rao-Blackwell theorem**. *Scholarpedia*, **3**(8):7039, 2008. 13

[36] El L Lehmann and Henry Scheffé. **Completeness, similar regions, and unbiased estimation: Part I**. *Sankhyā: the Indian Journal of Statistics*, pages 305–340, 1950. 13, 86

[37] Steven G Self and Kung-Yee Liang. **Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions**. *Journal of the American Statistical Association*, **82**(398):605–610, 1987. 14

[38] YP Mack and Murray Rosenblatt. **Multivariate k-nearest neighbor density estimates**. *Journal of Multivariate Analysis*, **9**(1):1–15, 1979. 15

[39] Rosenblatt M. **Remarks on some nonparametric estimates of a density function**. *Annals of Mathematical Statistics*, **27**:832–837, 1956. 16

[40] Parzen E. **On estimation of a probability density function and mode**. *Annals of*

Mathematical Statistics, **33**:1065–1076, 1962. 16

[41] Vassiliy A Epanechnikov. **Nonparametric estimation of a multivariate probability density**. *Theory of Probability & Its Applications*, **14**(1):153–158, 1969. 16

[42] Peristera P and Kostaki A. **An Evaluation of the Performance of Kernel Estimators for Graduating Mortality Data**. *Journal of Population Research*, **22**(2):185–197, 2008. 16

[43] Scaillet O. **Density estimation using inverse and reciprocal inverse Gaussian kernels**. *Nonparametric Statistics*, **16**(1-2):217–226, 2004. 16

[44] MC Jones and DF Signorini. **A comparison of higher-order bias kernel density estimators**. *Journal of the American Statistical Association*, **92**(439):1063–1073, 1997. 17, 18

[45] ParK B U and Marron J S. **Comparison of data-driven bandwidth selector**. *Journal of the American Statistical Society*, **85**(409):66–72, 1990. 18

[46] Park B U and Turlach B A. **Practical performance of several data driven bandwidth selectors (with discussion)**. *Computational Statistics*, **7**:251–270, 1992. 18

[47] Hall P, Sheather S J, Jones M C, and Marron J S. **On Optimal Data-Based Bandwidth Selection in Kernel Density Estimation**. *Biometrika*, **78**(2):263–269, 1991. 18

[48] Jones MC, Marron JS, and Sheather SJ. **A Brief Survey of Bandwidth Selection for Density Estimation**. *Journal of the American Statistical Association*, **91**(433):401–407, 1996. 18, 41

[49] Kanazawa Y. **Hellinger distance and Kullback-Leibler loss for the kernel density estimator**. *Statistics & Probability Letters*, **18**:315–321, 1993. 18

[50] Sam Efromovich. **Orthogonal series density estimation**. *Wiley Interdisciplinary Reviews: Computational statistics*, **2**(4):467–476, 2010. 18

[51] Geoffrey S Watson. **Density estimation by orthogonal series**. *The Annals of Mathematical Statistics*, pages 1496–1498, 1969. 18

[52] Hongwei Sun. **Mercer theorem for RKHS on noncompact sets**. *Journal of Complexity*, **21**(3):337–349, 2005. 20

[53] Mark Girolami. **Orthogonal series density estimation and the kernel eigenvalue problem**. *Neural Computation*, **14**(3):669–688, 2002. 20

[54] Boyd S and Vandenverghe L. *Convex Optimization*. Cambridge University Press, 2004. 20, 123

[55] Boyd S and Vandenberghe L. *Convex Optimization*. Cambridge University Press, 2009. 20

[56] Aluisio Pinheiro and Brani Vidakovic. **Estimating the square root of a density via compactly supported wavelets**. *Computational Statistics & Data Analysis*, **25**(4):399–415, 1997. 21, 22

[57] Adrian M Peter and Anand Rangarajan. **Maximum likelihood wavelet density estimation with applications to image and shape matching**. *Image Processing, IEEE Transactions on*, **17**(4):458–468, 2008. 21, 44, 45, 64

[58] Jian Cheng. *Estimation and Processing of Ensemble Average Propagator andIts Features in Diffusion MRI*. PhD thesis, University of Nice-Sophia Antipolis, 2012. 21, 22

[59] Ingrid Daubechies. **The wavelet transform, time-frequency localization and signal analysis**. *Information Theory, IEEE Transactions on*, **36**(5):961–1005, 1990. 22

[60] David L Donoho, Iain M Johnstone, Gérard Kerkyacharian, and Dominique Picard. **Density estimation by wavelet thresholding**. *The Annals of Statistics*, pages 508–539, 1996. 22

[61] Charles K Chui. **Wavelets: a tutorial in theory and applications**. *Wavelet Analysis and its Applications, San Diego, CA: Academic Press,— c1992, edited by Chui, Charles K.*, **1**, 1992. 22

[62] Carandoa D, Fraimanb R, and Groismana P. **Nonparametric likelihood based estimation for a multivariate Lipschitz density**. *Journal of Multivariate Analysis*, **100**(5):981–992, 2009. 24

[63] Coleman TP and Sarma SV. **A Computationally Efficient Method for Nonparametric Modeling of Neural Spiking Activity with Point Processes**. *Neural Computation*, **22**:2002–2030, 2010. 24

[64] Alfréd Rényi. **On measures of dependence**. *Acta mathematica hungarica*, **10**(3-4):441–451, 1959. 24, 25

[65] Berthold Schweizer and Edward F Wolff. **On nonparametric measures of dependence for random variables**. *The annals of statistics*, pages 879–885, 1981. 24, 25

[66] Marco Corazza and Elisa Scalco. **Verifying the Rényi Dependence Axioms for a Non-Linear Bivariate Comovement Index**. *University Ca'Foscari of Venice, Dept. of Economics Research Paper Series No*, **11**, 2015. 25

[67] WJ Hall. *On Characterizing Dependence in Joint Distributions*. University of North Carolina, Department of Statistics, 1967. 25

[68] Justin B Kinney and Gurinder S Atwal. **Equitability, mutual information, and the maximal information coefficient**. *Proceedings of the National Academy of Sciences*, **111**(9):3354–3359, 2014. 26, 56

[69] Thomas M. Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & sons, Inc., 1991. 26, 118

[70] David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. **Detecting novel associations in large data sets**. *science*, **334**(6062):1518–1524, 2011. 26, 56

[71] A Bhattachayya. **On a measure of divergence between two statistical population defined by their population dis-**

tributions. *Bulletin Calcutta Mathematical Society*, **35**:99–109, 1943. 28

[72] R Agarwal, Z Chen, and Sarma S V. **Nonparametric estimation of bandlimited probability density functions**. *arXiv:1503.06236v1, http://arxiv.org/pdf/1503.06236v1.pdf*, 2015. 34, 57, 58, 61, 62

[73] Gerhard J Woeginger. **Exact algorithms for NP-hard problems: A survey**. In *Combinatorial Optimization—Eureka, You Shrink!*, pages 185–207. Springer, 2003. 37

[74] Merz P and Freisleben B. **Greedy and Local Search Heuristics for Unconstrained Binary Quadratic Programming**. *Journal of Heuristics*, **8**:197–213, 2002. 37

[75] Taylor J. **First look - Gurobi Optimization**. Technical report, Decision Management Soluion, 2011. 37

[76] Raykar VC, Duraiswami R, and Zhao LH. **Fast computation of kernel estimators**. *Journal of Computational and Graphical Statistics*, **19**(1):205–20, 2010. 38, 50

[77] Hall P and Marron JS. **Choice Of Kernel Order In Density Estimation**. *Annals of Statistics*, **16**(1):161–73, 1987. 41

[78] Rae AIM. *Quantum Mechanics, 5th edition*. Taylor & Francis Group, 2008. 47

[79] Silverman B. **Algorithm AS 176: Kernel Density Estimation Using the Fast Fourier Transform**. *Applied Statistics*, **31**(1):93–97, 1982. 50

[80] H Skaug and D Tjøstheim. **Testing for Serial Independence using measures of distance between densities**. In Robinson P and Rosenblatt M, editors, *Athens Conference of Applied Probability and Time Series*. Springer, 1996. 55, 56

[81] Leandro Pardo Llorente. *Statistical inference based on divergence measures*, **185** of *Statistics, textbooks and monographs*. Chapman & Hall/CRC, Boca Raton, FL, 2006. 56

[82] Rahul Agarwal, Pierre Sacre, and Sridevi V Sarma. **Mutual Dependence: A Novel Method for Computing Dependencies Between Random Variables**. *arXiv preprint arXiv:1506.00673*, 2015. 58

[83] L De Lathauwer, B De Moor, J Vandewalle, and Blind Source Separation by Higher-Order. **Singular Value Decomposition**. In *Proc. EUSIPCO-94, Edinburgh, Scotland, UK*, **1**, pages 175–178, 1994. 62

[84] Howard Eichenbaum, Paul Dudchenko, Emma Wood, Matthew Shapiro, and Heikki Tanila. **The hippocampus, memory, and place cells: is it spatial memory or a memory space?** *Neuron*, **23**(2):209–226, 1999. 76

[85] Larry R Squire. **Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans.** *Psychological review*, **99**(2):195, 1992. 76

[86] Wendy A Suzuki, Earl K Miller, and Robert Desimone. **Object and place memory in the macaque entorhinal cortex**. *Journal of neurophysiology*, **78**(2):1062–1081, 1997. 76

[87] Edvard I Moser, Emilio Kropff, and May-Britt Moser. **Place cells, grid cells, and the brain's spatial representation system**. *Annu. Rev. Neurosci.*, **31**:69–89, 2008. 76, 85, 90

[88] Kass RE and Ventura V. **A spike train probability model**. *Neural Comput.*, **13**:1713–20, 2001. 77

[89] Sarma SV, Cheng ML, Eden U, Williams Z, Brown EN, and Eskandar E. **The effects of cues on neurons in the basal ganglia in Parkinson's disease.** *Front Integr Neurosci*, **6**(40), 2012. 77

[90] Santaniello S, Montgomery Jr EB, Gale JT, and Sarma SV. **Non-stationary discharge patterns in motor cortex under subthalamic nucleus deep brain stimulation.** *Front Itegr Neurosci*, **6**(35), 2012. 77

[91] Kahn K, Sheiber M, Thakor N, and Sarma SV. **Neuron Selection for Decoding Dexterous Finger Movements**. In *Proceedings of the 33rd IEEE EMBS Conference*, 2011. 77

[92] Gelman A, Carlin JB, Stern HS, and Rubin DB. *Bayesian Data Analysis*. CRC Press, 2003. 79

[93] Brown EN, Barbieri R, Ventura V, Kass RE, and Frank LM. **The time-rescaling theorem and its application to neural spike train data analysis.** *Neural Comput.*, **14**(2):325–46, 2002. 81, 82, 91

[94] Loren M Frank, Garrett B Stanley, and Emery N Brown. **Hippocampal plasticity across multiple days of exposure to novel environments**. *The Journal of neuroscience*, **24**(35):7681–7689, 2004. 90

[95] Robert U Muller and John L Kubie. **The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells**. *J Neurosci*, **7**(7):1951–1968, 1987. 90

[96] Jeff A Bilmes et al. **A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models**. *International Computer Science Institute*, **4**(510):126, 1998. 91

[97] Thomas G Dietterich. **Approximate statistical tests for comparing supervised classification learning algorithms**. *Neural computation*, **10**(7):1895–1923, 1998. 95

[98] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. **Data clustering: a review**. *ACM computing surveys (CSUR)*, **31**(3):264–323, 1999. 98

[99] Hazewinkel, M. *"Parseval equality", Encyclopedia of Mathematics*. Springer, 2001. 102

[100] Bertsekas DP. *Nonlinear Programming (Second ed.)*. Athena Scientific, 1999. 102

[101] I.M. Gelfand and S.V Fomin. *Calculus of Variations*. Dover Publ, 2000. 102

[102] Kobayashi H, Mark BL, and Turin W. *Probability, Random Processes, and Statistical Analysis*. Cambridge University Press, 2011. 108, 109

[103] Marcus Protzmann and Holger Boche. **CONVERGENCE ASPECTS OF**

BANDLIMITED SIGNALS. *Journal of ELECTRICAL ENGINEERING*, **52**(3-4):96–98, 2001. 112

[104] Simon Kochen, Charles Stone, et al. **A note on the Borel-Cantelli lemma**. *Illinois Journal of Mathematics*, **8**(2):248–251, 1964. 115

[105] Royden HL and Fitzpatrik PM. *Real Analysis, (4th Edition)*. Pearson, 2010. 117

[106] David Schmeidler. **Fatou's lemma in several dimensions**. *Proceedings of the American Mathematical Society*, **24**(2):300–306, 1970. 117

[107] Marks II RJ. *Introduction to Shannon Sampling and Interpolation Theory*. Springer-Verlag, New York,, 1991. 122

[108] Akaike H. **Block Toeplitz Matrix Inversion**. *SIAM J Appl Math*, **24**(2):234–41, 1973. 123

Vita

Rahul Agarwal received the MSE degree in Biomedical Engineering from Johns Hopkins University, Baltimore in 2011 and enrolled in the Biomedical Engineering PhD program during the same year. Earlier he recieved B. Tech degree in Electrical Engineering from Indian Institute of Technology (IIT), Kanpur in 2009. His research focus is on two seprate topics 1. developing novel methods to find structure in unstructured data and 2. understanding working of motor circuit in the brain. His papers have been published and are under consideration in reputed journals in Neuroscience, Computational Biology, Machine learning and statistics.

Starting in September 2015, Rahul will start as a staff scienist at St. Jude Medical corporation.